Toxoplasma gondii-mediated Host Cell Transcriptional Changes Lead to Metabolic Alterations Akin to the Warburg Effect

Lalitha Sridevi Sundaram

Girton College

August 2016

This dissertation is submitted for the degree of

Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the relevant Degree Committee. For more information on the word limits for the respective Degree Committee.

Dedicated to the memory of my mother, Prabha Sundaram, and my grandmother, Janaki Sundaram

Acknowledgements

I first would like to express immeasurable gratitude to my incredible, strong and loving family. Appa, Lakshmi and Ammamma – thank you for your patience and for always being there for me.

My heartfelt thanks to members of the Ajioka and Micklem labs, past and present: Russ Brown, Semil Choksi, Peter Davenport, Matthew Garrett, Krys Kelly, Bo Shiun Lai, Aysha Roohi and Orr Yarkoni. I'm so grateful to have gotten to know you – as lab-mates, and as friends. I've learnt so much from each and every one of you.

I'd also like to thank David Brown, Paul Dear, James Brown and Conrad Whittle, who have been an unlimited source of support and encouragement.

Finally, to Jim Ajioka and Gos Micklem – my supervisors and confidantes:
Your guidance, wisdom and friendship have truly been transformative.
I can never thank you enough.

Abstract

Toxoplasma gondii is an obligate intracellular parasite, that is able to infect any nucleated cell. An important global pathogen, *T. gondii* can cycle between primary and secondary hosts, thus enabling widespread penetrance. Within its intracellular niche – a membrane-bound parasitophorous vacuole – *T. gondii* is nevertheless able to subvert a variety of host cell processes to allow its continued survival and replication. This includes modulation of host signalling processes as well as the scavenging of nutrient macromolecules.

In recent years, microRNAs have emerged as important regulators of cellular processes including inflammation, tumorigenesis and metabolism, as well as development. It has become increasingly clear that this species of non-coding RNA is of great importance in 'fine tuning' many cellular responses. I hypothesise in this work that host cell miRNAs may be yet another means by which *T. gondii* manipulates its host upon infection.

Using high-throughput-sequencing, I examine host cell transcriptional responses to infection both at the mRNA and microRNA level, using two strains of *T. gondii* at a variety of Multiplicities of Infection over a time course of 43 hours. Through these transcriptional analyses I identify a number of dysregulated pathways common in tumorigenesis, and contemplate the hypothesis that *T. gondii* may be behaving as an 'intracellular tumour', subverting host cell metabolic processes to mimic a long-known feature of cancer metabolism – that of aerobic glycolysis (the Warburg effect) – in order to satisfy its own energetic and metabolic needs.

Table of Contents

List of Illustrations

VI. Effects of Parasite Infection on Host Cell Metabolism

VII. Discussion and Future Directions

List of Tables

Nomenclature

Throughout this work, I have attempted where possible to adhere to the conventions of nomenclature set out by the Mouse Genome Informatics Mouse Genomics Database (MGI, MGD) (Release 6.05).

Upon first mention, gene names are written out in full. Subsequently, they are shortened to their official MGI symbol. If the product being referred to is a gene or transcript, this symbol is rendered in italics. If it is a protein, the symbol is rendered in upper case letters.

Gene name: hexokinase 2
Gene symbol: *Hk2*
Protein: HK2

The exception is where long lists of genes are presented, and repetitive use of italics styling is unwieldy.

In certain cases, conventionally-used names for genes vary to a greater or lesser degree from the official symbol and can thus appear unfamiliar (e.g. *p53* as *Trp53* or *Hif1β* as *Arnt*). In these cases, I give the official full name and indicate the 'commonly-used' name upon first mention, after which I use the official nomenclature only.

A few special cases are gene families, such as NFkB, where the different members have very different official gene names and so I use the conventionally-used term as an umbrella, specifying individual names by official name or symbol where appropriate.

I. Introduction

1.1 *Toxoplasma gondii*

Toxoplasma gondii is an obligate protozoan parasite, capable of infecting a wide rage of hosts – virtually all warm-blooded animals, in fact. It is a coccidian apicomplexan, a member of a group which includes other intracellular parasites such as *Plasmodium* (the causative agent of malaria). As with most apicomplexans, *T. gondii* has a complex life cycle which includes both sexual and non-sexual phases. However, where many other coccidians have only the faecal-oral route for transmission, *T. gondii* has managed to widen its breadth of host infection by adapting to employ carnivorism and trans-placental infection, the latter of which has severe implications for human transmission. It is thought that *T. gondii* infects up to half of the world's population, exhibiting wide variation in different countries (16-40% in the United Kingdom compared to 50-80% infection in continental Europe, France having a notably high rate) (1). Despite this broad infection, the majority of human infections are asymptomatic, or at most characterised by mild, flu-like symptoms. However, in certain cases, the infection can result in toxoplasmosis and can have extremely severe, even fatal consequences. In immunocompromised patients, those infected with HIV or transplant recipients on immunosuppressive therapy for instance, toxoplasmic encephalitis (TE) can result (in fact, it is estimated that *T. gondii* infection kills 15-20% of all AIDS patients (2)). In women who experience the infection for the first time during pregnancy, the parasite is able to cross the placental border, and the foetus is left prone to hydrocephalus, intracranial calcification and chorioretinitis (3). Infection generally arises as a result of three things: ingestion of oocysts, ingestion of tissue cysts or congenital infection, acquired trans-placentally. This is better understood in the context of the parasite's life cycle.

1.1.2 Life Cycle

The sexual phase of the life cycle occurs only in the primary/definitive host, which are members of the cat family, but the asexual phase can take place in virtually any warm-blooded animal. *T. gondii* exists in several different forms including tachyzoite, bradyzoite, sporozoite, merozoite, micro- and macro-gametocyte. Each of these represents a stage in the parasite's development, and the first three are potential routes for infection. The tachyzoite (whose characteristic crescent-shape gives its name to the parasite: *toxo* meaning "arc" and *plasma* meaning "form") is the rapidly dividing asexual form of the parasite, able to actively penetrate cells of the infected animal. Bradyzoites on the other hand, divide slowly and reside in tissue cysts. Both of these forms occur in intermediate hosts. Infected cats, however, provide the appropriate environment for the sexual phase of the cycle, in their small intestine epithelial cells (4). This phase includes the production of gametes, and fertilisation. Maturation of the fertilised gamete also occurs here, and mature oocysts are then discharged (though they are at this stage unsporulated). Sporulation, which renders the oocysts infective, happens outside: this is the only form of the parasite that can survive outside its host.

Once sporulated, the oocysts – in water or feed contaminated with cat faeces – make their way into intermediate hosts (usually livestock or humans). Here, the parasite switches to the rapidly dividing form: this is the acute phase of infection, resulting in a concerted immune inflammatory response (hence the flu-like symptoms) and tissue necrosis. Once the parasite has actively penetrated its host's cells, it reverts to the bradyzoite form, and quiescent tissue cysts develop. Though these cysts can form in most visceral organs, they preferential develop in neural and muscular tissues, mostly being found in the brain, eyes, cardiac and skeletal muscle (5). Normally, these tissue cysts persist for lengthy periods of time (perhaps up to the lifetime of

the host, though this has not been conclusively shown), and, in immunocompetent hosts, should generally not permanently re-activate, though it may be that tissue cysts undergo multiple rounds of re-activation and re-encystment (6, 7). When these cysts are eaten by a naïve host, the parasite once again undergoes a phase-change and the bradyzoites revert to the rapidly-dividing tachyzoite phase, and the similar pathway of acute infection followed by encystation occurs in the predator (1). (Figure 1.1)

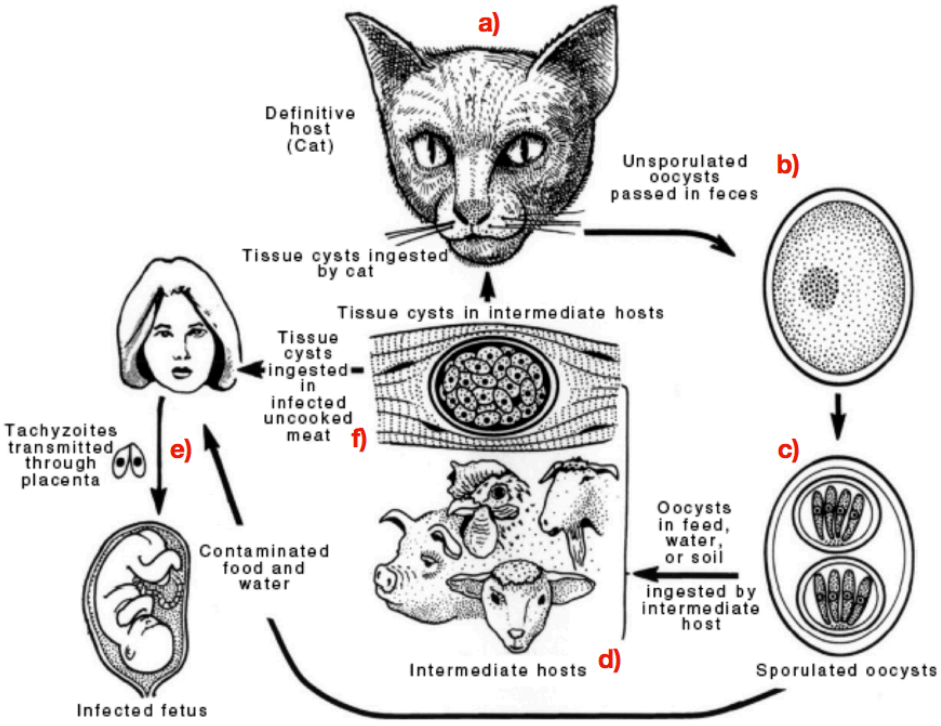


Figure 1.1 **Transmission and Life Cycle of *Toxoplasma gondii*.**
Toxoplasma gondii's definiteive host are members of the *Felidae*, where the sexual cycle occurs **(a)**. Following this, oocysts are shed **(b)** into the environment where, upon sporulation **(c)**, they may be eaten by intermediate hosts **(d)**. From here, there are two potential routes of further transmission: transplacental transmission of tachyzoites **(e)** and transformation into bradyzoites and subsequent ingestation of these tissue cysts **(f)** by other intermediate hosts or the definitive host. Adapted from (1)

1.1.3 Host Cell Effects

Since the parasite can only undergo genetic exchange when two strains infect a single cat, any recombinants from these events then go through selection in their secondary hosts, and this has had important consequences for the

population structure and likely evolution of the parasite. It appears that *Toxoplasma gondii* exists as four distinct clonal lineages in Europe and North America, within which strains are nearly identical (8, 9). These lineages exhibit marked differences in virulence and show distinct geographical distributions as well (9, 10). Most chronic infections in humans – and in animals eaten by humans – are caused by Type II strains. And, while infection with Type II strains rarely cause mortality in mice, in humans these are often associated with severe disease, such as congenital toxoplasmosis (10). Type I strains, though highly virulent in mice (with an LD100 of 1 parasite as compared to LD100 of $>10^3$ in Types II and III), are relatively rare in human disease. These differences in genotype and virulence have led to many 'classical' genetics experiments such as Quantitative Trait Locus mapping, to elucidate the basis for the phenotypic differences between strains (11). In addition to these lineages, a genetically diverse set of so-called 'atypical' South American isolates has also been identified. These often result in serious disease even when the hosts do not have a deficiency in immuncompetence and for this reason, their population structure is currently under investigation (12).

Following invasion of the host cell, *T. gondii* establishes its replicative niche through the formation of a parasitophorous vacuole (PV), from which host cell membrane components can be selectively excluded (13). Thus sheltered, the parasite can begin its replicative cycle through endodyogeny. As well as benefiting from the physical (albeit selectively permeable) barrier of the PV membrane (PVM), *T. gondii* additionally embarks upon a broad programme of host cell reprogramming, effectively 'hijacking' a number of endogenous host cell pathways and subverting cellular functions both to protect itself from host immune pathways but also (presumably) to render its environment more hospitable in terms of nutrient availability. The means by which *T. gondii* is able to do this are not completely known but has been shown in a few cases to be thanks to the discharge of parasite factors by the

parasite's secretory organelles: the rhoptries and dense granules.

Given the tractability of classical genetics in *T. gondii*, many such parasite effectors have been identified as a result of genetic crosses. For instance, ROP16 was identified as the parasite factor responsible for differential maintenance of STAT3/6 activation upon infection by strains of Types I and III (sustained activation) and Type II (transient). These differences were found to segregate into the F1 progeny of a IIxIII cross and subsequent QTL analysis pointed to the *Rop16* locus (11, 14). Given that many such parasite factors have been identified using (at least in part) these crosses, it is unsurprising that many of the now-known effectors are virulence-related (ROP18, ROP5, for instance(15, 16)) or strain specific in some other manner (GRA15 (17) or MAF1 (18), which mediates strain-specific mitochondrial association of PVs). Far fewer secreted factors have been identified that are not polymorphic among strains. Two such are GRA16, which was identified by computational prediction of putative secreted proteins followed by epitope tagging (19) and, more recently, a putative secreted protein MYR1 that underpins the induction of host cell myelocytomatosis oncogene (MYC), which was identified through a mutagenesis screen (20, 21).

While the exact mechanisms by which *T. gondii* is able to subvert host cell processes are not always known, there is a great deal more information on how these processes are modulated once that subversion has been initiated. The main areas of focus here have been the immune response and apoptosis (including but not limited to via NFkB).

At the centre of the host's innate immune response to infection are macrophages, whose protective actions are manifold. They are capable of responding directly to microbial products, resulting in the production of powerful reactive intermediates and can themselves engage in phagocytosis. Furthermore, they also represent an important link between innate and adaptive immunity: they can present antigenic peptides, and therefore

activate CD4+ and CD8+ T cells, which can then start clearing infection by eliminating cells in which parasites have taken refuge (22). The earliest events leading to recognition of *Toxoplasma* by the host are not as yet wholly clear, but it appears that a parasite-derived cyclophilin may bind to the chemokine (C-C motif) receptor 5 (CCR5). The downstream effect of this appears to be the setting in motion of the innate inflammatory response, triggered by the production of the cytokine interleukin 12 (IL12) from dendritic cells (DCs) (23). Another pathway responsible for these earliest events in the innate response to the parasite may be that triggered by Toll Like Receptors (TLRs), as disruption of downstream components of this pathway – myeloid differentiation primary response gene 88, MYD88 in particular – almost always lead to fatality in mice (24). This pathway again leads to the production of IL12, either through the NFkB branch, or via the MAPK one. In any case, following the recognition of the parasite by cells of the innate immune system, a programme of resistance is then begun.

The vigorous inflammatory response which is the result of the initial infection by the parasite is associated with high levels of pro-inflammatory cytokines. The major cytokine responsible for the control of the acute infection is IFNG, particularly given its role in inflammation and in activating macrophages to a microbicidal state. In fact, it was shown early on that after having been injected with anti-IFNG antibodies and infected with the parasite (with strain ME49, and infection being induced either intra-peritoneally or orally), most of the infected mice died during the acute phase of the infection (25). Moreover, hosts lacking IL12 – the cytokine responsible for stimulating the production of IFNG by NK cells and T cells – also succumb to unchecked tachyzoite proliferation (26). The production of IFNG by T- and NK- cells is mediated by transcription factors, STAT proteins and, importantly, members of the NFkB family (24).

The NFkB pathway is a critical host cell pathway that is modulated by a multitude of input signals and itself regulates a vast array of pathways. Most of the studied pathways that NFkB is involved in have been imunne-related or to do with apoptosis – both processes are highly relevant for *T. gondii* infection. A few of the stimuli of this pathway include mediators of infection such as lipopolysaccharide (LPS) as well as pro-inflammatory cytokines, as noted above, including TNF and IL1. The role of NFkB in *T. gondii* infection has been rather contentious, with reports both indicating activation and suppression at various different points of the pathway (27–29). This is rendered even more complex by the fact that several NFkB-dependent genes are themselves negative regulators of the pathway, and that among its myriad targeted genes are both 'pro-parasite' and 'pro-host' examples. For example, pro-inflammatory cytokines such as *Il12* and *Tnfa* are targets of NFkB's, but so are several anti-apoptotic factors, such as the Inhibitors of Apoptosis (IAPs). It is likely that host cell context and strain have a large role to play. Indeed, Type II GRA15 has been shown to activate the pathway in a strain-specific manner (17), but other mechanisms exploited by other Type strains cannot be ruled out. While the precise of role of NFkB in *T. gondii* infection has yet to be completely elucidated, one of its major effects – suppression of apoptosis – is definitely observed after parasite infection.

It is not surprising that apoptosis might be one of the mechanisms by which hosts defend themselves against pathogens, and this is observed in both bacterial and viral infection (30–32). Unlike necrosis, apoptosis is a highly-ordered process that follows many well-defined steps. It usually proceeds via one of two pathways of activation cascades, extrinsic and intrinsic (mitochondrial), depending on the stimulus. The mitochondrial mechanism has been related to pathogen invasion and the cascade initiated by a number of stimuli including infection, DNA damage or nutrient starvation. Following the mitochondrial pathway ultimately results in the formation of the

apoptosome, in a manner that is dependent on the release of cytochrome c, somatic (CYCS, commonly termed cytochrome *c*) from the mitochondria into the cytosol. Formation of the apoptosome ultimately results in the activation of the 'executioner' enzyme, caspase 3 (CASP3). The extrinsic pathway depends on the binding of a ligand such as tumor necrosis factor (TNF, usually called TNFα) or Fas ligand (TNF superfamily, member 6) (FASL) to a cognate receptor of the TNF "death family" to trigger activation of caspase 8 (CASP 8). Caspase 8 can then either directly or indirectly (through triggering CYCS release) activate CASP3 (33). *Toxoplasma gondii* has been shown to suppress host cell apoptosis when it was stimulated along either the extrinsic or intrinsic pathway (34).

Given that inappropriate implementation of an apoptotic programme could be disastrous for an organism, the control of apoptosis is very tightly managed. This is largely achieved through the balance of pro- and anti-apoptotic genes, specifically of the BCL-2 family (35), and it does appear that *T. gondii* can evade apoptosis by modulating that balance (29). However, this is not the complete story, as *T. gondii* has also been shown to prevent CYCS release and also interfere with the activity, recruitment or processing of the caspase enzymes themselves (36). The fact that apoptosis is at the convergence of a number of important host cell pathways effectively renders any of these a potential target for modulation by the parasite.

Merely evading host cell immune or apoptotic pathways is unlikely to be the entire story when it comes to *T. gondii* modulation of host cells, however. Being an obligate intracellular parasite, *T. gondii* depends upon its host to derive a number of nutrients (37) and there is evidence that host-to-PV trafficking can occur via host endocytotic pathways to serve this purpose (38). The parasite is a an auxotroph for a number of key metabolites, including purine (39), cholesterol (40), choline (41) and a number of essential amino acids (37, 39). Moreover, even when *T. gondii* is able to synthesis

metabolic building blocks itself, it may not always do so. While the mechanics of how this is achieved have not yet been fully elucidated, parasite modulation of the host cell environment in order to enable more fruitful scavenging must be considered.

1.2 MicroRNAs

The earliest hint of microRNAs (miRNAs) was found in *Caenorhabditis elegans*, through a number of experiments looking at larval development and the action of heterochronic genes. Genetic screens for new heterochronic genes identified *lin-4*, whose action was shown to be as a negative regulator of, among others, *lin-14*. This did not seem to be the action of a classical transcriptional-regulator, though, with many strands of evidence weaving together to present the case for RNA-RNA interactions. While levels of the target gene's transcripts remained constant, LIN-14 protein levels decreased in a temporal manner, a decrease for which the gene's 3' Un-Translated Region (UTR) was required. Investigation into the regulator itself, *lin-4*, revealed that it was unlikely to encode a protein: neither start nor stop codons were identified and nonsense mutations did not alter the rescuing power of the *lin-4* product. Instead, it seemed that the two products of the gene – a barely-detectable 61nt transcript and a more abundant 22nt transcript – were the active species as RNAs, a notion that was strengthened by the fact that both transcripts had the potential to base pair directly with the target mRNA. Secondary structure for the larger transcript was also proposed as being a stem-loop, a structure now known to be characteristic of miRNA precursors. Because of this gene's involvement in developmental timing, their products were termed small temporal RNAs (stRNAs) (42).

A second heterochronic gene that appeared to encode a functional regulatory RNA with potential binding to its target's 3' UTR was discovered

seven years later, in the laboratory of Gary Ruvkun (43). The discovery of a second stRNA, *let-7*, was intriguing enough but this study also raised the idea that the scope of regulation by stRNAs might be wider than previously thought. *Let-7* appeared to have several targets, which included *lin-14*. Thus, it became likely that these stRNAs were part of a much more complex regulatory network: they were each able to downregulate the expression of multiple genes, and target genes themselves were likely to have multiple stRNAs as regulators. That same year, a larger scale study of *let-7* conservation revealed the presence of perfect homologues in the genomes of *Drosophila melanogaster* and humans. Moreover, *let-7* expression was demonstrated to occur in 14 other species, in many of which it appeared to be temporally expressed (44).

The true extent of this new kind of gene regulation only really became apparent thanks to the emergence and elucidation of RNA interference (RNAi) and its mechanisms (45, 46). It was while examining silencing RNA intermediates (siRNA) in lysates of *D. melanogaster* embryos injected with long dsRNA RNAi precursors that Elbashir et al (47) noticed the potential presence of endogenous small RNAs the same size as the siRNA products they had wanted to characterise. These endogenous small RNAs, along with 37 candidates from HeLa cells, were characterised later that year (48) and found to be expressed as 21-22nt RNA transcripts, to apparently come from precursors that folded into hairpins and, largely, to be conserved across organisms. Consecutive papers in that same issue of Science (49, 50) characterised further small RNAs in nematodes: It was now clear that these regulatory RNAs did indeed represent an "abundant class" (49). Beyond what had already been elucidated with *lin-4* and *let-7*, however, the roles of these new regulators was unclear and could not necessarily be presumed to involve development: it was thus agreed that the term microRNA would be used to refer to these genes.

1.2.1 Structure, Biogenesis, Mode of Action

As with the discovery of miRNAs themselves, elucidation of their biogenesis owes much to work done on pathways involved in RNAi. In early 2001, the Hannon laboratory identified the RNAse III family nuclease, Dicer, as the enzyme responsible for cleaving dsRNA silencing precursors into the active 21-22nt molecules involved in silencing (51). Later that year, Hutvágner et al showed that this same enzyme was also the one responsible for processing of the stem-loop precursor (pre-miRNA) transcripts into the mature (21-22nt) form (52). Here, following transfection of Dicer-targeted siRNA, HeLa cells were examined for *let-7* expression. In contrast to un-transfected or control siRNA-transfected cells, HeLa cells in which Dicer had been silenced exhibited an absence of mature *let-7* along with an accumulation of the larger (~72nt) precursor.

Unlike tRNAs and snRNAs, miRNA genes are transcribed by RNA Polymerase II (53, 54), as 'conventional' long transcripts, including the stabilising 5' cap and a polyadenylated tail. Once transcribed, the pri-miRNAs undergo a first round of processing within the nucleus by an enzyme complex termed the 'microprocessor' (55, 56). This complex recognises the pri-miRNA through binding of the protein DiGeorge syndrome critical region gene 8 (DGCR8, also known as PASHA) (57) which then allows the RNAse III family member DROSHA (58) to process the transcript into a 60-70nt hairpin structure, the pre-miRNA. Exportin 5 then mediates transport of this precursor, in a Ran-GTP-dependent manner (59, 60) to the cytoplasm, where the pre-miRNA is processed by a second protein complex, again containing an RNAse III family-member (DICER) and a dsRNA binding protein (Trans-activation Response RNA Binding Protein, TRBP). Cleavage by DICER removes the loop, resulting in the production of a short (~22nt) RNA duplex molecule (consisting of the miRNA and miRNA* arms of the stem) with 2-3nt

overhangs. The mature (miRNA) strand of the duplex is then preferentially incorporated into the miRNA-induced silencing complex (miRISC) (61) which mediates the repression of target mRNAs.

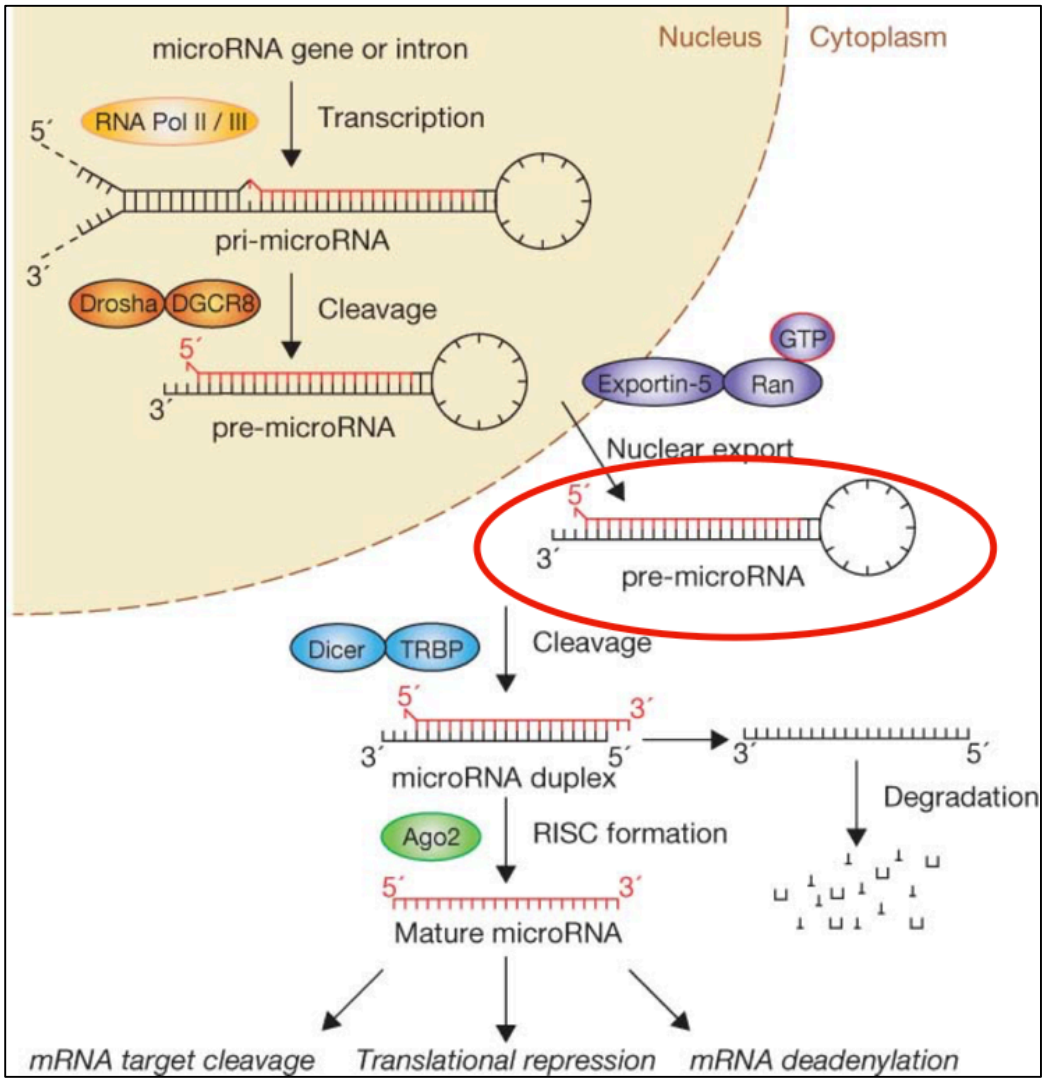


Figure 1.2 **Canonical microRNA biogenesis, with the characteristic stem-loop structure providing the basis for computational analyses of miRNA-seq data.**

After transcription, the nascent pri-miRNA is cleaved by the microprocessor complex (which includes Drosha and DGCR8) into the pre-miRNA (circled in red). The pre-miRNA is the stage at which most miRNA-seq experiments assay for discovery and profiling of miRNAs, given its distrinctive stem-look structure. My own analyses in subsequent chapters search for this characteristic pattern. From here, it is exported to the cytoplasm where its loop is cleaved off by the Dicer/TRBP complex. One half (previously termed star strand) is degraded, leaving behind the mature miRNA. The RISC complex, which includes Ago2, then mediates the main mechanism of gene regulation. Adapted from (62).

The concept of regulator non-coding RNAs is not new. As early as 1969, Britten and Davidson set forward a model of gene regulation in the journal Science, that rested on non-repetitive, non-coding sequences as master-controllers (63). This theory did not gain wide acceptance at the time however, as the term 'junk DNA' began to grow in use to refer to non-coding DNA (64). Despite this somewhat pejorative term, other types of functional yet non-coding RNA have been well-characterised since Britten and Robinson's publication, and several of these non-coding RNAs have been revelead to be vast classes with complex functions.
Some features of each of these is described below:

piRNAs (PIWI-interacting RNAs): Like miRNAs, piRNAs are small in size, approximately 24-31nt and also mediate post-transcriptional regulation. Unlike miRNAs, however, this class of small non-coding RNAs performs its functions by associating in a complex with the PIWI subfamily of Argonaute proteins (rather than the Ago subfamily) and appear to function primarily through epigenetic silencing of, for example transposable elements, rather than transcript silencing (65). PIWI proteins were first identified in *Dropsophila*, in terms of germline development and as such, the bulk of piRNA research has focused on regulation in this context. That being said, recent studies have shown piRNA expression in somatic tissues as well, with dysregulation being a feature in, for example, several types of tumour (66).

snRNAs (small nuclear RNAs): Small nuclear RNAs were first discovered in the late 1960s, and have been shown since then to play an important role in the maturation and processing of mRNAs. They are often divided into two classes, Sm- and Lsm, based on both specific sequence features as well as which protein partners are required for their function. Sm-class snRNAs are transcribed by RNA polymerase II and exported to the cytoplasm where they

are processed into stem-loop structures and where the ribonucleoprotein Sm core particle is formed (this process requires a number of protein cofactors such as the Survivor Motor Neuron complex). Upon re-importation into the nucleus, most of the snRNA snRNAs form part of the spliceosome and thus are key to the correct progression of intron-removal. Unlike other Sm-class snRNAs, U7 snRNA appears to mediate histone processing (67).
The second class of snRNAs, Lsm, comprise U6 and $U6_{atac}$ genes, which are transcribed by RNA polymerase III and, unlike the Sm-class, do not leave the nucleus. They too aid in the splicing process, and are essential for correct removal of introns (68).

snoRNAs (small nucleolar RNAs): Small nucleolar RNAs represent another size range of non-coding RNAs, from 60-300nt. An abudant and diverse class, snoRNAs can be subgrouped based on conserved sequence motifs: C/D or H/ACA. They were initially characterised in terms of rRNA processing, by through 2'-O-methylation or pseudouridination. However, they are increasingly being found to function in telomere synthesis, modifications of tRNAs or even mRNAs (67). As with snRNAs, snoRNAs operate within ribonucleoprotein complexes, the assembly of which is a highly directed process.

lncRNAs (long noncoding RNAs): As with the other main classes of ncRNA, lncRNAs have been identified in several animal species, including mouse, zebrafish and humans, as well as in plants such as *Arabidopsis* and maize. In the realm of parasites, lncRNAs have been identified in Plasmodium Toxoplasma and Neospora. Despite a few well-studied examples (such as the *Xist* gene which has a large role in X chromosome inactivation via dosage compensation), classifications of lncRNAs (in terms of mechanism or secondary structure, for instance) appears difficult to generalise, other than

they are over 200nt in length, transcribed by RNA polymerase II, and can be capped and polyadenylated, and even spliced. As such, one popular method of classification and naming has relied on position in reference to the genomic context. Here, lncRNAs are classed based on whether they are: intronic without overlapping exons; intergenic; bidirectional (in relation to the promoter of the nearest protein-coding gene); antisense (overlapping exons of a protein-coding gene but in the antisense direction); or sense-overlapping, sometimes called transcribed pseudogenes which, while overlapping a protein-coding gene's exon, do not produce a protein (69). These RNAs have been shown to be involved in a multitide of cellular regulatory processes including transcription, RNA degradation, translation, chromatin remodelling and splicing (69, 70). Moreover, lncRNAs have been found to function in processes involved in immune modulation, development, chromatin modulation and have also ben shown to be dysregulated in disease states. This has prompted Ulitsky and Bartel to comment, of the intergenic subset though this observation could probably be extended to other classes of lncRNAs, that they are defined "more by what they are not than by what they are" (71). High-throughput inhibition screens via CRISPRi are now underway, to be able to probe lncRNA function at scale and early results indicate involvement in regulation of cell growth in a cell-type specific manner (72).

1.2.2 Methods of Discovery

1.2.2.1 The 'Early Days'

Though the first miRNAs (*lin-4* and *let-7*, as above but also *bantam*, a *Drosophila* miRNA that emerged from a transposon-based gain-of-function screen (73)), were discovered by classic forward genetics, such screens have not been very successful in identifying new miRNA genes. This is partially a logistical problem: the small size of the molecules means that mutations in

these loci are relatively rare, and, given the often (at least partially) redundant roles of miRNAs, successful mutants are difficult to score (74).

That the mechanisms of siRNA production involved the same machinery as the processing of miRNAs proved to be very useful when it came to the identification of new candidates. While characterising siRNAs from exogenously-introduced dsRNA, Elbashir et al (2001) found that these fragments contained 3' hydroxyls and a 5' terminal phosphates, features that they exploited by specifically adding sequencing adaptors to either end of a size-fractionated library with T4 RNA Ligase (which catalyzes the formation of phosphodiester bonds between nucleic acid termini with these modifications). The ability of this method to also enrich for endogenous miRNAs meant that it could be used to concatamerise, clone, and sequence these genes specifically (47).

This process proved fruitful in the early years of miRNA discovery, with several groups employing it to characterise novel miRNAs. In the same 2001 issue of Science mentioned previously, three groups expanded the number of known miRNA genes from two (*let-7* and *lin-4*) to 93 (48–50), in humans, nematodes and *Drosophila*.[1] Cloning of material from specific organ and tissue fractions, cell lines or developmentally-staged oraganisms further increased this number, with, for example, 34 novel miRNAs being characterised (some with highly tissue-specific expression) from mouse heart, livers and a variety of brain fractions (75), and 40 novel miRNAs from rat neuronal tissue (76). A large-scale 'atlas' of mammalian miRNAs was compiled in 2007 (77), with miRNAs cloned and sequenced from 250 distinct mouse, human and rat tissues.

Despite the wide success of this method, it quickly became clear that cloning was approaching its limits in terms of miRNA discovery. miRNAs that

[1] This number refers only to the number of novel miRNAs discovered through cloning. In the cited works, the authors also employed phylogenetic comparisons to demonstrate expression of several more new miRNA genes in the named species as well as others.

are expressed at low overall levels or are expressed in a tissue- or developmentally-specific fashion might be overlooked. In fact, as early as 2003, it was asserted that the upper limit for novel miRNA discovery through cloning had been reached (78). The aforementioned large-scale atlas of mammalian miRNAs, though providing a comprehensive (at the time) and well-characterised compendium of miRNAs, only yielded a relatively meagre 33 novel miRNAs discovered through cloning. Moreover, these methods of cloning were very labour-intensive and the (conventional) sequencing that followed often proved expensive.

These considerations all indicated that alternative strategies for miRNA discovery needed to be developed.

1.2.2.2 Computational

Computational methods for identifying novel miRNA genes have largely been based on the premises of conservation among related organisms and structure (and, often, conserved structure). Though the sequence of steps might vary, most early methods either looked for conserved non-coding sequences and then folded them *in silico* or vice versa, identifying regions that were likely to fold according to certain rules and then comparing these to known miRNAs. The two first such miRNA-finding computational methods embody these strategies.

miRscan (79), developed in the Bartel laboratory, identified candidate stem-loops in the *C. elegans* genome and then searched for conservation in *C. briggsae*. Thus filtered, candidate stem-loops were then themselves scanned and scored for attributes gleaned from analysis of the 50 then-known *C. elegans* miRNAs, such as the extent of base-pairing within the stem, or the symmetry of bulges. When this approach was 'trained', using the *C. elegans* and *C. briggsae* genomes, half the known *C. elegans* miRNAs were found. Applying miRscan to the human genome using mouse and, subsequently,

(Taki)fugu rubripes (pufferfish) genomes for phylogenetic comparisons, identified 74 of the known 109 human miRNAs. Interestingly, the authors at the time argued that their method could be used to assign an upper bound to likely miRNA genes in the human genome. While their calculated figure was of 255, the current number of human miRNA genes in the current version miRBase (80) (version 21) stands at 2,588[2], perhaps speaking to the inherent limitations of using a limited set of known miRNAs and conservation between distantly-related organisms as definitive criteria.

Another, possibly more effective computational method for identifying new miRNAs was miRseeker. Instead of beginning with a search for hairpins, this program instead looks for all conserved, non-coding sequence alignments between two or more genomic samples. Identified orthologous fragments are *then* folded *in silico*, and the resultant hairpin loop is scored on the basis of the quality of its folding. Structures that have passed this filter then are scored and ranked on the basis of whether or not they follow canonical miRNA evolution patterns in the divergence of their nucleotides (as described in (74)). Of the 24 then-known *D. melanogaster* miRNAs, 18 were in the top 124 candidates identified by the program. Moreover, miRseeker predicted 48 novel miRNAs, 24 of which were experimentally verified through Northern blotting.

1.2.3 Methods of Verification / Profiling

Such computational methods generate lists of putative miRNAs whose expression requires validation through experimental means, if the putative miRNA is to be included in miRBase. This can be achieved through a variety of means, from single-miRNA profiling via Northern blots to more recent highly-scalable deep-sequencing platforms. These techniques can also be used

[2] Though miRBAse has recently done away with the notion of *miRNAs so this number may be inflated.

to examine the potential differential expression of miRNAs across experimental conditions of interest.

Northern Blotting

As might be expected, given the well-established nature of the technique, Northern blots have been widely used to validate predicted miRNA expression profiles. However, their disadvantages are also obvious: a (comparatively) large amount of starting material is necessary to obtain useful signals, largely due to relatively low-hybridisation capacities of conventional oligonucleotide probes. These probes themselves need to be radioactively labelled – a practice that, increasingly, is falling out of favour due to safety concerns. Additionally, traditional methods of cross-linking RNA to membranes prior to probing (such as through exposure to UV light) are often suboptimal when considering molecules as short as miRNAs (81). A number of variations to the 'classic' Northern blot protocol have been developed to address these specific concerns. Improvements include the use of digoxygenin-labelled probes instead of ^{32}P-labelled ones (82), Locked- Nucleic Acid (LNA)-modified oligonucleotide probes (which resulted in a tenfold improvement in sensitivity (83)), or a modified cross-linking buffering system, using a water-soluble carbodiimide, 1-ethyl-3- (3-dimethylaminopropyl), to immobilise the RNA to the membrane (81).

More recently, a method has been devised that combines the benefits of all three modifications to the standard Northern blot protocol (84). Kim et al's 'LED' protocol (LNA, EDC, DIG) was shown to be capable of detecting as low as 0.01–0.025 fmoles of synthetic miRNA with an exposure period of one minute (a great improvement from overnight or longer exposures typically required by isotopic methods). Nevertheless, despite these improvements, Northern blots for miRNAs are still time-consuming and laborious, given that only a few transcripts can be analysed at once. Moreover, the high cost of

LNA probes makes the method more suited to validation or profiling of just a few candidate miRNAs rather than to large-scale expression studies.

Primer Extension, q-RTPCR

Quantitative Real-Time PCR has been widely used as a verification measure for microarrays, and several groups have attempted to employ this method for the quantitation of miRNA expression as well. As with many of the other methods of miRNA analysis however, the short length of miRNAs raises significant problems. For a start, the length of primers used in qPCR is typically as long as the miRNAs themselves, and the criteria normally observed in primer selection for conventional qPCR to ensure specificity are difficult to adhere to. To address this, several modifications to primer design have been developed. Chen et al (2005) use stem-loop primers to increase the specificity of annealing, by hindering the adherence of non-specific (or pre-miRNA) sequences and increasing the stability through base-stacking. This is followed by the addition of a TaqMan® miRNA-specific fluorescent/quencher probe and Taq polymerase. As the reaction proceeds along the template strand, the polymerase's 5' – 3' exonuclease activity separates the fluorescent moiety of the probe from its quencher, yielding a measurable signal. Using this modified assay, the authors were able to distinguish between a miRNA precursor and the mature form, at a sensitivity of over 2000-fold (85).

The TaqMan® assay has been extended for use as parallelised reactions, sold commercially as Array Cards, which consist of a 384-well plate with TaqMan® probes to individual miRNA targets per well). Similarly, Raymond et al (2005) have used LNA primers to increase the binding affinity of primers to target miRNAs. In 22 of the 30 miRNA assays that they conducted, they found that the inclusion of LNA bases in the primers either "significantly enhanced" or was "absolutely required" for primer extension and miRNA profiling (86).

Improved sensitivity afforded by stem-loop or LNA-containing primers aside, the use of miRNA-specific probes can often prove expensive. Varkonyi-Gasic et al further modified this assay to enable the use of universal reverse primers, instead of the costly LNA or fluorescent TaqMan® probes. The authors found that, at 35 PCR cycles, 20pg of starting material were sufficient to produce an appreciable signal (though a greater number of cycles resulted in a modest amount of non-specific amplification) (87).

Microarrays

The advantages of using microarrays for global transcriptional analyses have been well-documented, especially in the field of parasite-host research (88). Their relatively low cost, the ease of procuring starting material, as well as the existence of numerous thoroughly-tested programs for data post-processing (such as normalisation, clustering of genes and differential expression profiling) make arrays an attractive choice for large scale RNA profiling. Indeed, many of these advantages remain when arrays are applied to the analysis of small RNAs, but this class of molecule also presents its own challenges. Traditionally, the design of probes for microarrays has been based on selecting regions of genes with as much unique sequence as possible. In this manner, the likelihood of mRNA derived from a different gene hybridising to the probe can be minimised, even when other regions of genes may be highly similar. The small size of miRNAs greatly limits the scope of probe selection, however and thus the risk of mis-hybridisation is severe. This problem is exacerbated due to existence of miRNA families, whose members often differ by only a few nucleotides. Biases inherent in array hybridisation only add to these concerns: if the region of a miRNA that distinguishes it from a family member contains a run of nucloeotides that preferentially hybridise to the array, then not only is transcriptional information about the 'correct' miRNA

lost, but the expression profile of the closely-related, 'incorrect' miRNA is amplified, and may then be falsely labelled as being significant.

The comparison between a TaqMan® qPCR array and a single conventional oligonucleotide array revealed that the false-discovery rate of differentially-expressed miRNAs was significantly higher using the microarrays than the q-RT-PCR (when measured as significantly-different expression levels between replicates on each platform). Moreover, overall correlation between the two platforms was also low (89).

The challenges of using such short sequences as microarray probes has been addressed by a number of techniques aiming to increase the specificity of hybridisation. Wang et al's array platform for Agilent (90), for example, uses hairpin sequences as probes, effectively blocking the hybridisation of longer matching target RNAs (pre-miRNAs, or degradation products from mRNAs, for example). Exiqon too have sought to improve hybridisation specificity by altering their probe-design. Here, LNA-containing oligonucleotides are used as probes, which results in a specific increase in melting temperature for perfectly-matched probe-target heteroduplexes. In fact, Castoldi et al found that even a single mismatch was enough to reduce the Tm of hybrids and result in enough destabilisation to appreciably lower the signal produced by such cross-hybridisation (91). Ambion's (now defunct) MirVana arrays introduced spacers into the probe sequence, thus increasing the effective hybridisation length.

Another kind of array modification to enable the accurate profiling of miRNAs has been to employ on-chip enzymatic reactions. In Nelson et al's RAKE (RNA-primed Array-based Klenow Enzyme) method, arrays are spotted with oligonucleotides consisting of a miRNA-specific antisense probe, followed by a short stretch of thymidines and a 5' spacer (which is consistent across all spots). After hybridisation with target RNA, the slide is treated with exonuclease, to degrade single-stranded spots (unbound by miRNA). The

klenow fragment of DNA polymerase then uses the bound miRNA as a primer and the spotted oligonucleotide as a template to then introduce biotin-conjugated dATPs only to the miRNA-bound spots. Fluorescent labelling of these biotins then results in a measurable signal for the direct assessment of the miRNA targets. Advantages of this method include the fact that it includes neither reverse transcription nor amplification steps (both of which might introduce bias) and probe hybridisation is based on the 3' end of the miRNA: for closely-related miRNAs, this is the area where nucleotide divergence is most often seen. In fact, the authors show that for six tested miRNA paralog pairs, RAKE distinguished between the pairs better even than Northern blotting (92).

Next-Generation Sequencing

In the past decade or so, the advent of Next-Generation or High-Throughput sequencing has profoundly altered the way that we think about genome-scale experiments, and the area of miRNA study is no different. This method boasts a number of advantages over those described in this chapter, and as such it is what I have employed for this work. I discuss its various instantiations as well as their strengths and limitations in **Chapter 3**.

1.3 Thesis Overview

MicroRNAs provide a potent means of altering a number of cellular processes and their potential is only now beginning to be understood. Thus, I wanted to explore what their impact was on host-parasite infection. To do this, I employed a variety of next-generation sequencing approaches to explore both the modulation of host miRNAs but also the possible existence of parasite-encoded ones. RNASeq lends itself well to large-scale probing of host-cell interactions so I extended this study to look at mRNA alterations as well. The transcriptional observations (supported by the literature) that infection with

T. gondii results in a remodelling of host metabolic systems from numerous signalling angles then led me to probe the issue of aerobic glycolysis more directly, looking at protein levels of key enzymes in different host genetic backgrounds as well as using this metabolic remodelling as a (very putative) drug target.

II. Materials and Methods

2.1 Cell Culture - General

2.1.1 Cell Culture Medium

All cell and parasite culture manipulations were performed in a Containment Level 2 Facility. Only fixed or lysed cellular material was removed from the room, except for a) receipt from or transfer to another laboratory, b) transfer to the -80 °C refrigerator for cryopreservation. In both these cases, live material was double-contained.

Routine cell culture was performed in high glucose Dulbecco's Modified Eagle's Medium (Sigma-Aldrich) containing 4500 mg/L glucose, L-glutamine, sodium pyruvate, and sodium bicarbonate. To this medium was added Foetal Bovine Serum (Sigma-Aldrich, 10%), Penicillin/Streptomycin (Sigma-Aldrich 100U/ml-100µg/ml), and 2mM L-glutamine (PAA), and the preparation was disposed of no later than two weeks of preparation. Thereafter, this supplemented medium is referred to as HG-DMEM.

2.1.2 Mycoplasma Testing

Cell and parasite cultures were tested upon first receipt into the laboratory for contamination by *Mycoplasma spp.* This was done either using the Lonza's MycoAlert™ Mycoplasma Detection Kit (manufacturer's instructions) or by sending a sample of cell culture supernatant to the Wellcome Trust – Medical Research Council's Cambridge Stem Cell Institute for analysis. All live cells and parasites were subsequently tested when an infection (or any aberrant cellular behaviour) was alerted by any other user of the Containment Level 2 Facility.

2.2 Host Cell Culture

2.2.1 Host Cell Strains

Host cells for routine parasite maintenance, for the RNASeq experiments and for the lactate assays were NIH/3T3 Mouse Embryonic Fibroblasts bought from DSMZ (Catalogue Number ACC-59). Host cells used for the Western Blots in Chapter 6 were HIF1A-KO and HIF1A-WT Mouse Embryonic Fibroblasts, a kind gift from Dr Ira Blader, SUNY Buffalo, New York.

2.2.2 Host Cell Freezing

Cells were seeded into 75 cm^3 flasks two to three days before freezing. When cell monolayers achieved ~60% confluence, they were deemed suitable for cryopreservation. Cells were rinsed twice with PBS and detached using 1x Trypsin-EDTA (placed in the incubator to aid detachment). Following detachment, HG-DMEM was added to the cell suspension. Cells were pelleted by centrifugation (300 x g) and then resuspended in HG-DMEM for counting with an Improved Neubauer haemocytometer. Following counting, the cells were centrifuged, rinsed with PBS, and then re-centrifuged. The final cell pellet was then resuspended in enough cold freezing medium (FBS/10%DMSO) to ensure a final concentration of 1 million cells per mL. Following resuspension, cells in freezing medium were aliquoted into pre-cooled cryovials, at a volume of 1 mL (1 million cells) per vial. These vials were placed in a Nalgene freezing container overnight at -80°C before subsequent storage in liquid nitrogen. Passage number was increased upon cell freezing.

2.2.3 Host Cell Thawing

Cryovials (stored as described in **2.1.2**) were rapidly thawed in a 37° C water bath. Contents of the cryovials were then added dropwise to pre-warmed HG-DMEM. Following centrifugation at 300 x g, cell pellets were resuspended in

HG-DMEM and the resulting suspension transferred to cell culture flasks. Cell culture medium in these flasks was replaced the following day.

2.2.4 Host Cell Routine Maintenance (Subculture/Passage)

For routine maintenance, cells were grown in 75 cm^3 flasks and were not allowed to exceed 80% confluence at any time. Monolayers of appropriate 60-70% confluence were washed twice with 1X PBS to remove traces of serum, and trypsin-EDTA was then added. Cell culture vessels were then returned to the incubator (37 °C, 5% CO_2) to promote detachment, for 1-2 minutes. Detachment of cells from the vessel surface was then confirmed microscopically, and HG-DMEM added in order to halt the action of the trypsin-EDTA.

Following collection of cells by centrifugation (300 x g), cell pellets were resuspended in HG-DMEM. Aliquots of the cell suspension were then applied to a each chamber of an Improved Neubauer haemocytometer and the number of cells/mL determined. The appropriate volume of cell suspension was then added to a fresh flask along with enough HG-DMEM to make up the correct volume for the size of the flask being used. Passage number was increased by one every time a culture was passaged. Cell lines were not allowed to exceed a passage number of 25.

2.3 Parasite Culture

2.3.1 Parasite Strains

Type II Strain: ME49 B7 Clone or PTG-GFP

Type I Strain: RH 88 or RH-GFP

All parasite strains were a kind gift from Dr David Sibley, Washington University School of Medicine.

2.3.2 Routine Maintenance

Parasite maintenance and propagation followed a pattern of lysis and re-inoculation, whereby as soon as an infected host cell monolayer was fully lysed, an appropriate volume of the lysate was added to an uninfected flask of ~70% confluent host cells.

2.3.3 Parasite Harvest

In order to maximize parasite yields, ~70% confluent host cell monolayers were inoculated 42 hours prior to parasite harvest such that the majority of host cells were observed (microscopically) had been infected, by at least 4 parasites and yet remained un-lysed.

Infected monolayers were then rinsed twice with 1X PBS to remove extracellular (and thus potentially dead) parasites. HG-DMEM was then added, and the monolayers were scraped using a rubber policeman. The infected host cells were then mechanically lysed by passing first through a 25G needle and then through a 27G needle, 5 times each. The lysate was filtered through a 3 μM polycarbonate filter membrane to remove host debris and the parasites from the filtrate were then collected by centrifugation at 1000 x g. Parasite pellets were resuspended in HG-DMEM. Aliquots of the parasite suspension were then applied to each chamber of an Improved Neubauer haemocytometer and the number of cells/mL determined. The appropriate volume of parasite suspension was then added to fresh HG-DMEM/10%FBS, which was then used to replace the medium in 70% confluent flasks or wells of host cells. Where a particular Multiplicity of Infection (MOI) was specified for an experiment, an additional three host cells were seeded and counted, in order to better determine an accurate MOI.

2.3.4 Parasite Freezing

NIH/3T3 cells seeded in 25 cm^3 flasks were infected with parasites when they were ~70% confluent. When heavily infected (though not lysed), the freezing protocol was begun. Cells were washed twice with PBS and detached with Trypsin-EDTA and the detachment was stopped with pre-chilled 50%FBS/HG-DMEM and chilled on ice for two minutes. Pre-chilled 20%/DMSO was then added to bring the volume up to 2 mL total per starting 25 cm^3 flask. This suspension was transferred to pre-chilled cryovials (1 mL per vial) and transferred to a Nalgene freezing container. The freezing container was placed at -80 °C overnight and the vials then transferred to liquid nitrogen.

2.4 Small RNA Library Preparation

Given that much of the preparatory work in **Chapter 3** was concerned with refining and optimizing Illumina's in-house protocol (Solexa microRNA Sample Prep Protocol, v1.4B), the details of library preparation are covered in **3.2**.

2.5 Bioinformatic Methods

Given that the bioinformatics methods used for these studies often required modification and/or evaluation, the are discussed in the relevant chapters.

2.6 RNA Extraction (for Chapters 5 and 6)

RNA extractions were done according to the DirectZol instructions. All steps were conducted using RNAse-free labware and on surfaces treated with RNAseZap (Sigma Aldrich). Briefly, for each well in which cells had been grown/infected: Each well was washed twice with PBS, treated with 950 µL Tri Reagent (Sigma) and mixed well by pipetting. After five minutes, the mixture was transferred to a tube and centrifuged at 12,000 x g for one

minute. The supernatant was then transferred to a fresh tube and 950 µL Ethanol was added. A Zymo-Spin™ IIC Column was placed in a collection tube and loaded with 700 µL of the mixture. After centrifugation for one minute, the collection was emptied and the procedure repeated until all the mixture had been spun. The column was then placed into a new column and rinsed twice via centrifugation for one minute with 700 µL Direct-zol™ RNA PreWash solution. A wash step followed, using 700 µL of RNA Wash Buffer and a one minute centrifugation. After a final centrifugation step to ensure complete removal of the wash buffer, the column was eluted in 25 µL DNAse/RNAse-free water. Sample drying is discussed in **Chapter 5.**

2.7 Lactate Assays

Lactate assays were performed using Abcam's Colorimetric/Fluorometric L-Lactate Assay Kit (ab65330), according to the manufacturer's instructions. Briefly, host cells were grown in 25 cm^2 flasks until they were ~70% confluent. At this point they were infected or simply had their medium changed and returned to the incubator. Extracellular medium was removed, centrifuged (1000 x g) to remove any potential parasite or cellular debris, diluted by a factor of two and placed on ice. Lactate standards were reconstituted and the assay was applied to standards and samples at the same time in black 96-well flat-bottomed plates (Greiner). After incubation, the plates were visualized using a Labtech FLUOstar Omega plate reader.

2.8 Western Blot

After infection (or growth) for the suitable time, ice-cold RIPA Buffer (Sigma Aldrich) with cOmplete™, Mini, EDTA-free Protease Inhibitor Cocktail (Roche) was added to 25 cm^2 flasks. These were placed on ice for five minutes after which the lysate was scraped using a rubber policeman. Lysates were transferred to pre-chilled tubes and stored until further use.

On the day of the western blot, the protein samples were clarified by centrifugation at 8000 x g for ten minutes. Protein concentrations were determined for each sample using a BCA Protein Assay (Thermo Scientific Pierce) according to the manufacturer's instructions.

20 μg of each sample was added to an equal volume of 2X Laemmli sample buffer (Bio-Rad) with 30X Dithiothreitol (DTT) and the mixture boiled at 95 °C for five minutes. Following this, the lysate mixture was loaded into the wells of a 4-20% SDS-PAGE gel (Bio-Rad) and the gel was run at 150 V for one hour in 1X TGS.

After the gel was run, it was soaked in Transfer Buffer (25 mL Tris/190 mM/20% Methanol) for fifteen minutes and then assembled into a transfer sandwich with a methanol-wetted membrane. Proteins were then transferred for 2.5h at 35 V. Transfer quality was checked with a Ponceau S stain following which the membrane was incubated with the primary antibody in 5% Milk/TBST, overnight at 4 °C. The blot was then rinsed in TBST three times for ten minutes and then incubated with HRP-conjugated secondary antibody for one hour at room temperature. Following another three rinses as before, the blot was then incubated with ECL Plus Western Blotting Detection Reagents A and B at a 1:1 ratio (GE Healthcare) for one minute. After removal of the chemiluminescent substrate, the membrane was developed and visualized onto X-ray film (Fuji). [All steps of Western Blot apart from lysate preparation were done with Bo Shiun Lai]

2.9 Fluorescence Microscopy

Cells were seeded (and, if necessary, then infected) onto sterilized, gelatine-coated coverslips. Though NIH/3T3 and HeLa is an adherent cell lines and would thus not usually need an additional substrate on the coverslips, the number of washes involved in the fixation process and the fact that monolayers are rendered somewhat fragile when infected meant that it was

necessary to coat sterilized coverslips with gelatine to ensure secure adherence. This was done by first sterilizing 22 mm^2 glass coverslips in 70% ethanol (six times per coverslip), allowed to dry briefly on lint-free tissue paper and then rinsed twice in separate containers of 1X PBS. After blotting the edges on lint-free tissue paper, the coverslips were then placed in the wells of 6-well plates. A 2% gelatin solution (Sigma Aldrich) was added to each well and the plates returned to the incubator for 30 minutes. After this, the gelatin was aspirated and the plates allowed to dry.

After growth, infection and/or treatment with methyl jasmonate (Sigma Aldrich), conditioned medium was removed from the wells and replaced with ice cold 4% paraformaldehyde (Pierce). After 15 minutes, this was aspirated and disposed of according to local safety guidelines. The wells were rinsed three times with pre-chilled PBS and either covered with PBS for storage at 4 °C or mounted immediately. Coverslips were mounted onto glass slides (Superfrost) using Vectashield Antifade Mounting Medium with DAPI (Vector Labs) and the edges sealed using a clear nail varnish said not to autofluoresce (Kai-Fai Leung, personal communication; Boots Natural Collection).

Images were captured using the AF6000 system with a Leica DM6000B upright fluorescence microscope and prepared for publication here using Fiji is just ImageJ, Fiji (93).

III. Identification of Putative Novel microRNAs

3.1 Introduction

3.1.1. High-Throughput-Sequencing / Next-Generation Sequencing

The advent of high-throughput sequencing in the past ten years has changed many facets of genomic biology. The particular technical features (read lengths almost exactly the correct size, for example) and the emergence of the technology at a time when microRNAs too were just being discovered has meant that the application of NGS has been very useful in the field of miRNA research.

Several methods of high-throughput sequencing have been developed and, though the methodologies may vary, the common features of all these platforms include those of parallelisation, speed of data generation and (relatively) low cost. Along with the ability to quickly produce vast amounts of sequence data, the fact that they determine sequence by 'counting' a signal produced by base-incorporation, rather than by interpreting an often-noisy hybridisation signal means that they can also be used for Digital Gene Expression (DGE). Put simply: counting the number of sequenced transcripts can thus be used as a direct measure of the expression of that transcript and so, these methods provide a robust platform both to discover new miRNAs and to determine their individual expression levels.

The three 'market leaders' in high throughput sequencing are Roche's 454 system, ABI, with their SOLiD platform and Illumina's Genome Analyzer, each with application-specific advantages and disadvantages (94).

The 454 method, first commercialised in 2005 (95) essentially involves a bead-based parallelisation of pyrosequencing. Here, sheared sequence fragments are attached to beads such that a single fragment is bound per bead. Following amplification and denaturation of the bound fragments (resulting in a bead covered with identical template sequences) the beads are

then deposited into single wells, along with a light-generating enzyme mix. Sequential 'waves' of nucleotides are then applied to the wells: if a nucleotide is incorporated into the following position of the single-stranded, bead-immobilised template, a stoichiometric number of photons is emitted. For each wave of applied nucleotides, the amount of light generated in each well is recorded and translated into sequence. The proportionality of light generated to the number of identical nucleotides incorporated into the template means that a single cycle is not necessarily limited to having extended the template has been extended by only one nucleotide, and thus, reads from 454 sequencing technologies can be much longer than those produced by other high-throughput technologies (with read lengths of ~400bp, although far fewer reads will be generated. However, this advantage is also a drawback: the quantitation of emitted light intensities has been shown to falter when presented with for longer stretches of incorporated nucleotides. As a result, insertions or deletions (indels) of one or more nucleotides are the most common sequencing errors of this type of technology (96, 97). The long length of produced reads, along with 454's propensity to produce indels rather than substitution errors (which are, arguably, easier to account for when the length of the read to be aligned is limiting (96, 98) mean that the application of 454 to miRNA discovery and profiling – where accuracy, and depth of coverage are valued over read-length – would not be using the technology to its best advantage.

ABI's SOLiD platform for DGE is also bead-based but proceeds somewhat differently, employing a sequencing-by-ligation technique with di-base interrogation. Here, adaptors are ligated to the ends of sequence fragments and made to bind to beads, where they are amplified via emulsion PCR. Sequencing primers complementary to the bound adaptor are added, along with ligase and four fluorescently-labelled 'di-bases'. These di-base constructs consist of two nucleotides in a specific order followed by three

degenerate bases, ending with one of four fluorescent modifications, such that each of the four fluorescent colours itself corresponds to one of four possible dinucleotide combinations. Each position along the extending template sequence is thus tested for the *n*th and the *n+1*th base, and the correct di-base (with the appropriate sequence of two nucleotides) is ligated and the fluorescence read. Following cleavage of the fluorescent moiety, the steps are repeated for the sequence immediately following the three degenerate (uncleaved) bases, yielding base-calls for sequence positions *n*, *n+1* (from the first cycle), *n+5* and *n+6* (from the second cycle). After five such cycles, the primers are stripped and sequences re-probed with progressively shorter primers, so that eventually, each nucleotide position along the template sequence has been interrogated twice, with a different di-base combination. After the desired number of rounds, the resultant 'gapped' sequences are intercalated in so-called 'colour-space' and matched to a matrix of each of the possible di-nucleotide combinations to determine the full sequence.

While this technology has advantages when it comes to distinguishing sequencing errors from natural variation (and is thus very useful when it comes to applications such as SNP calling, where these will have specific colour variations in the sequencing process), the fact that it involves not only per-cycle fluorescence cleavage steps but also several rounds of sequencing with different primers greatly increases the time taken for a run (99). Additionally, its use of colour-space, rather than the more familiar nucleotide- or base-space as a sequencing 'alphabet' means that end-users are usually unable to interpret 'raw data' themselves and must rely on SOLiD-specific tools for alignment.

The sequencing technology implemented by Illumina, has, arguably, proven to be the most-widely adopted one, especially in the United Kingdom (according to a crowd-sourced map of sequencing facilities (100).

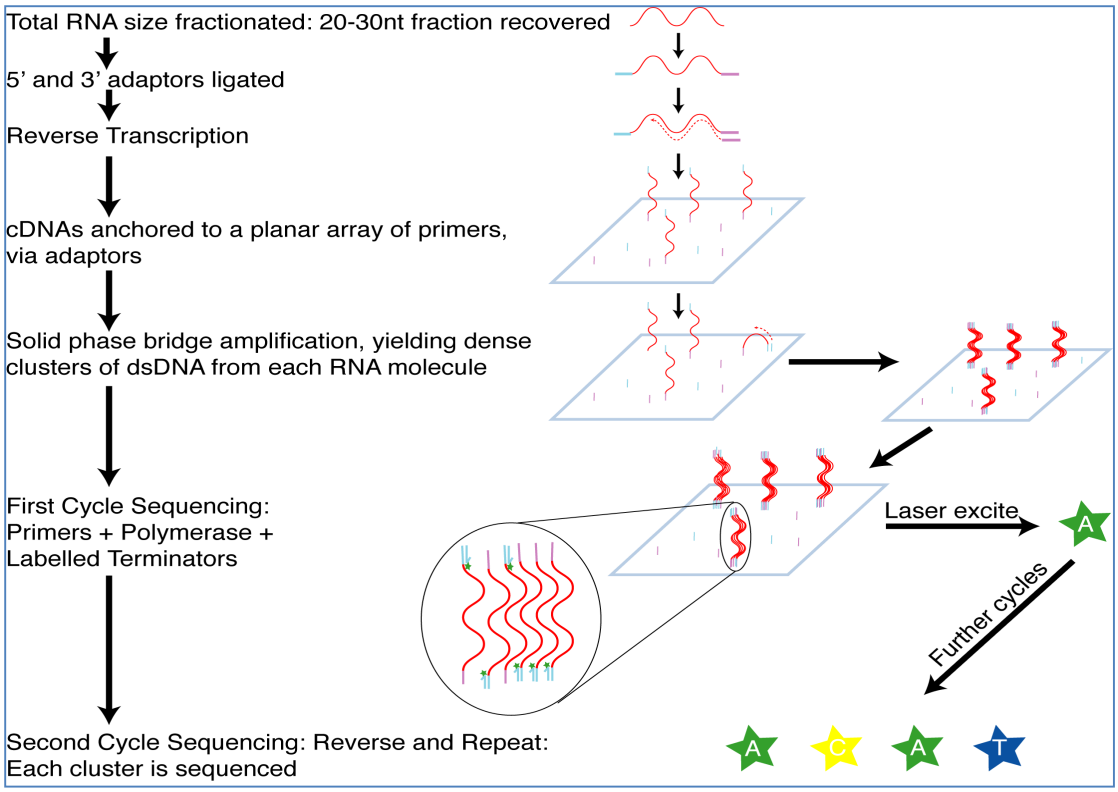


Figure 3.1. A schematic representation of the Illumina sequencing process.

The main steps involved in the Illumina sequencing-by-synthesis process are shown here, with RNA fragments depicted in red; 3' and 5' adaptors in blue and violet. Individual nucleotides are represented as coloured stars.

The first steps of the chemistry are very similar to other high-throughput sequencing platforms, where genomic or transcriptomic sequence fragments of interest are ligated to adaptors. In the case of miRNA analyses, the adaptors are conceived such that they will only ligate to molecules possessing the characteristic 3' OH and 5' phosphate termini. Instead of being immobilised and amplified on beads, however, these templates are covalently attached to a slide where solid bridge amplification takes place. Much like Sanger sequencing, this technology relies on the use of dye-labelled terminator nucleotides but, importantly, these terminators are reversible (Reversible Terminator bases, RT-bases (87)). Instead of having to electrophorese and sort each fluorescently-terminated fragment by size (as in Sanger sequencing) after each cycle of Illumina's RT-base addition, the fluorescence is read, and

the terminator is cleaved (leaving behind a now-unmodified nucleotide), allowing the reaction to proceed anew for the following base (Figure 3.1).

The process of having to incorporate, read and cleave, for every extended base, renders the sequencing reaction quite slow (Illumina's own estimates for the Genome Analyzer were of ~48 hours for a 36-cycle run (102), though this figure has vastly improved with newer sequencers. Another result of the sequencing-by-synthesis chemistry of this method is that the length of resultant reads is quite short: At the technology's inception, read lengths were only 27nt in length (103) though both improvements in the technology and the use of paired-end or mate-pairing methods now yield far longer reads, which make the technology especially useful for RNASeq (**Chapter 6**). Regardless, such short read lengths are not an issue when it comes to miRNA discovery and characterisation.

3.1.2 Bioinformatic Challenges

As with other platforms of DGE, Illumina's Genome Analyzer technology is challenging on many levels. Though the sequencing itself has been greatly facilitated since the days of directional cloning, methods of library preparation are still complex multi-stage processes and, at least in the early days of implementation, little documentation was available. Once the sequencing has been performed, these experimental challenges are replaced by bioinformatic ones. The sheer volume of returned sequence means that traditional algorithms for interpretation are no longer feasible, and instead, novel tools need to be employed. While several such packages have emerged, in the early years they tended to be application- or platform-specific, having often been developed 'in-house' to tackle a single lab's specific concerns (Krys Kelly, personal communication, Rory Stark and Kevin Howe, personal communication). When in-house methods are published as generalised tools, they may often simplify the process without fully describing what assumptions

are being made and whether these are applicable to the problem at hand (104, 105). Indeed, many publications that employ DGE often describe their bioinformatic methods in a very cursory manner, without presenting arguments for each of the steps taken. Moreover, despite the relatively recent history of DGE, technologies and platforms are being changed and improved extremely quickly, meaning that software tools, unless actively maintained and updated for each hardware release on each platform, may not be able to keep pace.

3.1.3 Novel miRNAs in *Toxoplasma gondii*

To-date, only a handful of studies have sought to identify and characterise parasite miRNAs directly. Braun et al were the first to do so, by deep-sequencing purified parasites of Type I RH, Type II Prugniaud and Type III CTG strains. They identified 14 putative miRNA families and performed northern blots on these. Seven families gave a positive signal, though this was often with 'double banding'. The authors controlled for the possibility that these may be host miRNAs by running northern blots on uninfected host RNA as well, and found no visible signal (106). However, none of these miRNAs have to-date been included in miRBase, though a comment on the website provides a reason: A comment by the miRBase creator, Sam Griffiths-Jones in February 2013 states, "We do have a few non-animal, non-plant miRNAs that have been published and submitted to us. The sets we don't have were not submitted or it was agreed with authors not to deposit them. There is some controversy in the community about these" (107). He further goes on to cite a review that challenges the veracity of published putative miRNAs from a number of species, those identified by Braun et al in *Toxoplasma gondii* included (108). The main challenge from the review's author is that the miRNAs identified by Braun et all either cannot be localised to the genome, lack star sequences, or can be shown to be fragments

of other RNAs, such as a perfect match to *T. gondii* 18S rRNA. This is a risk of their analysis procedure, where they excluded rRNAs from their read set before performing alignments, rather than looking for rRNA fragments *post hoc*. Additionally, the fact that they employed a fairly low Phred cut-off score of 10, perhaps indicates that the libraries may have been of overall low-quality. Given the group's success in validating at least some of the candidates by northern blot does not of course preclude at least some of their putative miRNAs from being non-canonical, or 'miRNA-like' RNA species.

The next study of *Toxoplasma gondii* miRNAs was performed in 2012 by Wang et al (109), where miRNAs from RH and ME49 were compared. Interestingly, their parasites were purified from mice (rather than propagated in cell culture). Curiously, their cited method for parasite purification is for separating bradyzoites cysts from the brains of infected mice – it is not a usual method for tachyzoite isolation from mice, which would usually be done through intraperitoneal lavage of infected mice (Uas Müller, Stuart Woods, personal communication). Nevertheless, the authors maintain that "tachyzoites were purified" by this method (109). Nevertheless, their analysis revealed 356 novel putative miRNAs, of which 17 were found to be related to 2 metazoan miRNA families (which are unnamed in the manuscript). That being said, a quick look at the sequences that are presented as novel putative miRNAs show some as being entirely composed of GT dinucleotides. Of course, some annotated miRNAs within the current instantiation of miRBase are similar, such as mmu-miR-466i, but that does not make them any more 'real'. The authors find that seven of their putative miRNAs are supported by the Braun et al (110).

Xu et al also used tachyzoites from live mice to perform their analyses, and they identified tachyzoites from intraperitoneal lavage of infected Kumning mice. The authors were able to identify 54 putative miRNAs from five strains of *T. gondii*. In an unusual choice, they did not also collect and

profile intraperitoneal lavage samples of uninfected mice, nor did they align their sequence reads to the mouse genome (111).

The most recent analysis of putative *T. gondii* miRNAs is a wholly computational effort which holds as its central hypothesis the possibility that *T. gondii* export hairpins to the host where they would be processed and would act on host cell targets. Saçar et al first extracted all possible hairpins from the genome of *T. gondii* (ME49) and then used a machine learning approach to compare these to mouse, rat and Chinese hamster mature miRNAs, as a scoring filter. The resultant high-scoring hairpins were then checked against publicly-available *T. gondii* transcriptomic data. Apart from the fact that there is no evidence of pre-/pri-miRNA export form the parasite, this is nonetheless a compelling idea, one that has been proposed before (112). One of the limitations of Saçar et al's study, however, is that their step of aligning their putative *T. gondii*-derived miRNAs to parasite transcriptomic data (to confirm expression) did not take into account any of the characteristic patterns of miRNA read alignment. Moreover, though they were looking specifically at miRNAs ostensibly of parasite origin, the transcriptomic datasets they were using would almost certainly contain host material as well. The fact that they cross map parasite-derived rodent-like miRNAs to the parasite transcriptome grown in human cells (and vice versa) might have been an attempt to 'correct' for this, but no mention is made in the text, and the reciprocal alignment (which would have been performed had this been a concern) was not performed.

Several of these studies highlight the need to look carefully at the genomic context in which miRNA precursors are encoded, and how that impinges on expected patterns of alignment and folding. At its most basic, a particular profile is expected, with a large stack of reads aligning preferentially to one of the stems of the pre-miRNA hairpins but with a few reads aligning

to the star strand and an even lesser proportion crossing the hairpin region. This is illustrated in Figure 3.2.

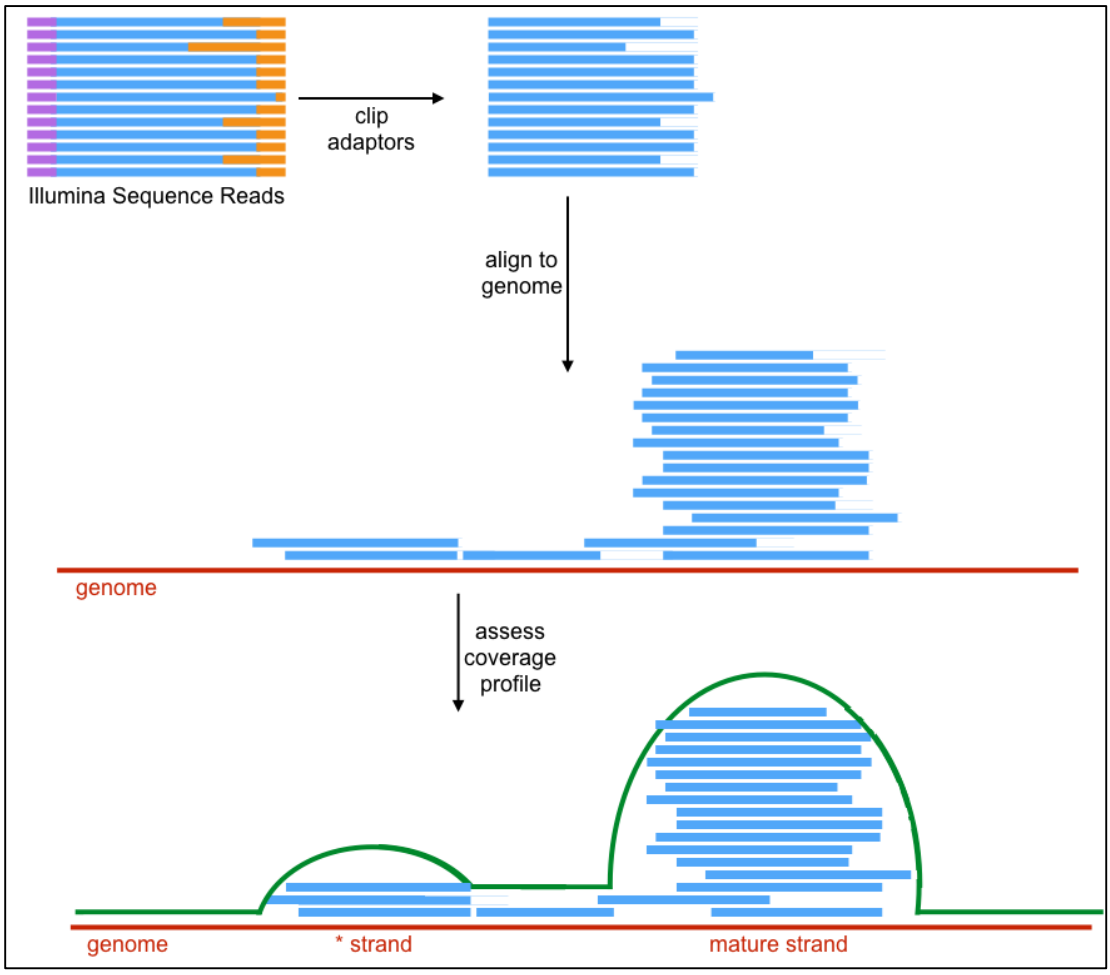


Figure 3.2. A schematic representation of the steps involved in NGS analysis.
After sequencing reads (blue) have had their adaptors (orange and purple) clipped off, they must be aligned to the genome (red). If they truly do originate from a miRNA-encoding locus, a characteristic profile is expected (green).

Indeed, this is one of the criteria that miRDeep2 (113), a programme developed by Friedländer et al to identify novel miRNAs from sequencing data, uses. First, potential precursors are excised from the genome, based on the presence of alignment 'stacks' corresponding to potential mature, star and loop sequences. These potential precursors are then folded, using RNAfold and the secondary structures of these are evaluated and scored, based on the

likelihood of forming an unbifurcated hairpin loop (as compared to the appropriate randfold calculations). The potential precursors are also scored based on whether they contain a seed sequence similar to known miRNAs from a closely-related species and on whether they have been detected in more than one of the input samples. A list of known inputted miRBase miRNAs is also subjected to the same process as a control procedure, in order to then estimate the sensitivity of the programme (which is then fed back into the prediction scores for the putative novel miRNAs. As such, each potential miRNA receives a final combined miRDeep2 score from which the user can, if desired, select the criteria which are most likely to represent truly novel miRNAs given the particular experimental conditions.

In this chapter I will discuss the application of Illumina's GA technology – from sample library preparations to analysis of the resultant data to predict novel miRNAs – when applied to NIH/3T3 mouse embryonic fibroblasts mock-treated or infected for 24 hours with *Toxoplasma gondii* tachyzoites. I discuss several alternative methods of handling each step of the process, paying particular attention to some of the pitfalls that may be faced when dealing with these types of data.

3.2 Methodology

NIH/3T3 were infected with Type II, ME49 tachyzoites, at a Multiplicity of Infection (MOI) of 5:1 or left uninfected for 24h, after which total RNA was isolated using a TRI Reagent / column purification method. This material was collected by Dr Nadia El-Guendy in the lab of Dr Anthony Sinai at the University of Kentucky.

3.2.1 MicroRNA Library Preparation

Small RNA libraries were prepared from these RNA samples according to the Illumina-recommended protocol (version 1.4B). Given that this section deals

largely with probing several aspects of this protocol to better understand the mechanisms of library preparation, I include it here (rather than in **2.4**). Briefly, the recommended protocol is as follows:

1) Total RNA is electrophoresed on a 15% Tris-Borate-Urea (TBU) polyacrylamide gel alongside 10bp DNA ladders.

2) The (invisible) gel slice corresponding to the miRNA fraction (20-30nt) is then excised, pulverised and the RNA eluted for 4h in NaCl, following which the 5' RNA Adaptor is ligated (for 6h).

3) The resultant preparation is electrophoresed again, on a 15% TBU polyacrylamide gel.

4) The (invisible) gel slice corresponding to RNAs of size 40-60nt is then excised, eluted for 4h in NaCl and ligated to the 3' RNA Adaptor.

5) The full RNA construct (70-90nt) is electrophoresed on a 10% TBU gel, the corresponding (invisible) band excised, and the RNA eluted for 4h.

6) Following RT-PCR using a tailed primer, the resultant amplified DNA product is then electrophoresed on a 6% Tris-Borate-EDTA polyacrylamide gel.

7) The (now-visible) band is excised, eluted and sequenced.

Given the lengthy and multi-step process involved in library preparation, I decided to examine several of the steps in detail and optimise them for efficiency if possible.

Ladder Calibration

Because the first few steps of the Illumina library preparation protocol involved the excision of bands that would be not be visible, it was important to ensure that the DNA ladders that were to be used as guides were well-calibrated and migrated at rates similar to the RNA samples. To this end, I ran 10 and 25bp DNA ladders on a hand-cast 15% TBU polyacrylamide gel, alongside DNA and RNA standards of known sizes. I also included Drosophila

melanogaster total RNA, since that organism's 2S ribosomal RNA is of similar size to miRNAs (30nt) (114) and could be used as additional RNA marker for ladder calibration. It appeared that there was some discrepancy between the migration rates of RNA and DNA (Figure 3.3, lanes 3-5) with the DNA samples (including the ladder) running marginally faster. This has been reported for formaldehyde-containing agarose gels (115) but may also, as we have found, have a small effect in polyacrylamide gels. This slight discrepancy was taken into account in all gel excision steps where a DNA ladder was used as a guide for an RNA sample.

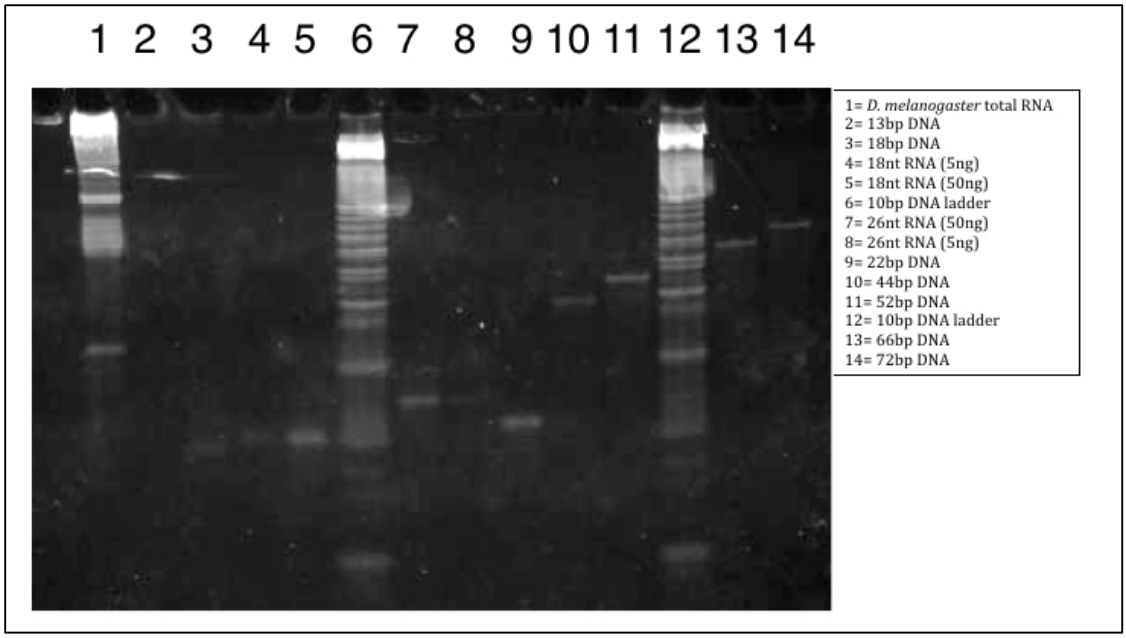


Figure 3.3 Analysis of migration rates between RNA and DNA reveals a slightly faster migration for DNA samples.

A variety of RNA and DNA samples were run on a polyacrylamide gel to assess the migration of these two different species, and whether this might have any implication for downstream library preparation methods. The results from Lanes 3 (18bp DNA), 4 (18nt RNA, low concentration) and 5 (18nt RNA, high concentration) indicate that DNA appears to migrate slightly faster than RNA counterparts in polyacrylamide gels. The remaining lanes were run to ensure that any discrepancies arising from migration speeds were indeed a result of the DNA/RNA difference and not a 'systemic' difference between nucleic acid fragments and the corresponding bands on the reference ladder.

Ligation

The unit definition for T4 RNA ligase is given as catalysing 1 nanomole of 5´-[^{32}P]rA$_{16}$ into a phosphatase-resistant form in 30 minutes at 37°C and so, the ten units of enzyme specified by the protocol coupled with incubations of six hours seemed excessive. I verified this by performing the ligation reaction with an RNA standard for either the recommended six hours or for one hour, and found that the shorter incubation period was sufficient, as shown in Figure 3.4. Indeed, more recent versions of the Illumina small RNA library preparation protocol perform ligations for one hour per adaptor (116).

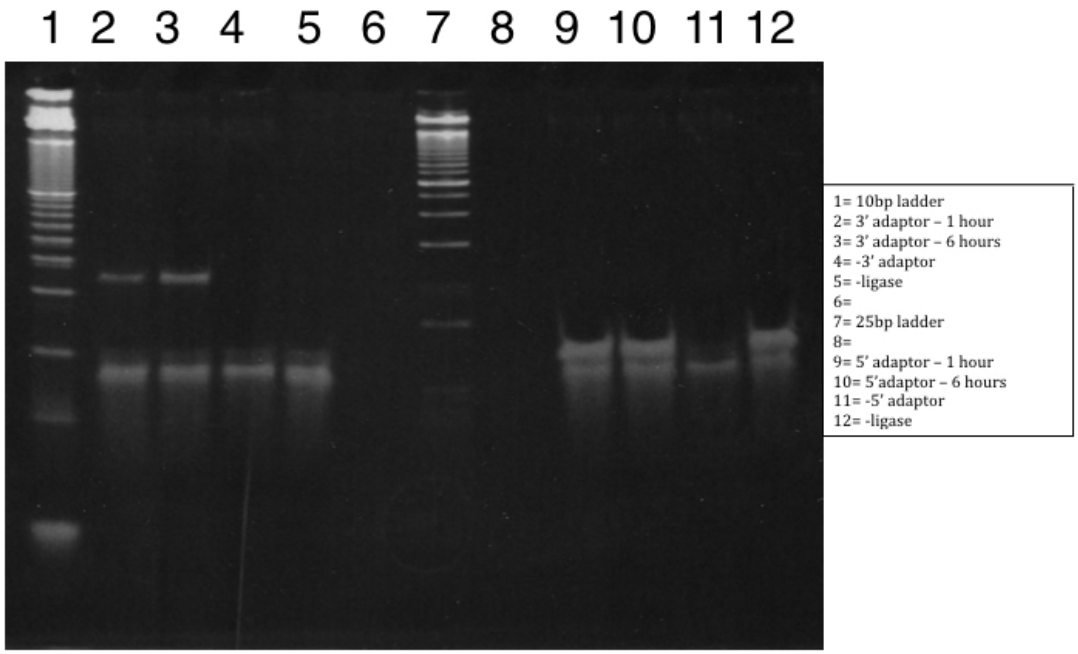


Figure 3.4. **Analysis of different ligation times shows that a one-hour incubation is sufficient to yield sufficient ligation.**

Pairs of each ligateable or not a RNA standard (3' and 5' adaptor) were incubated with T4 RNA ligase for either one or six hours (the latter being recommended in the Illumina protocol). Lanes 2 and 3 show that a one hour incubation is sufficient to ligate 3' adaptors together. Controls were performed with the 5' adaptor, lanes 9 and 10, which will not ligate due to the absence of a phosphate at the 5' end and an OH group at the 3' end.

Samples

The version of the microRNA library protocol that I followed recommended the inclusion of a co-precipitant carrier after the elution of the cDNA band, in

order to facilitate the identification of pellet. The manufacturers of Pellet Paint® note in their instructions that, while the carrier itself absorbs in the UV range used for nucleic acid quantitation, a batch-specific correction factor can be applied to offset its contribution to absorbance. This proved not to be the case with my samples, however, resulting in extremely erratic and inconsistent Nanodrop readings. As an alternative quantitation, I therefore performed gel electrophoresis of 1µL of each sample alongside various dilutions of DNA ladders. Though a number of cDNA libraries were prepared, only four were used for downstream analyses. The samples in Figure 3.5 labelled in red are the ones that were used for subsequent downstream analyses. FB1.1 and FB2.1 are replicates of uninfected cells while FB1.3 and FB2.3 are replicates of infected cells.

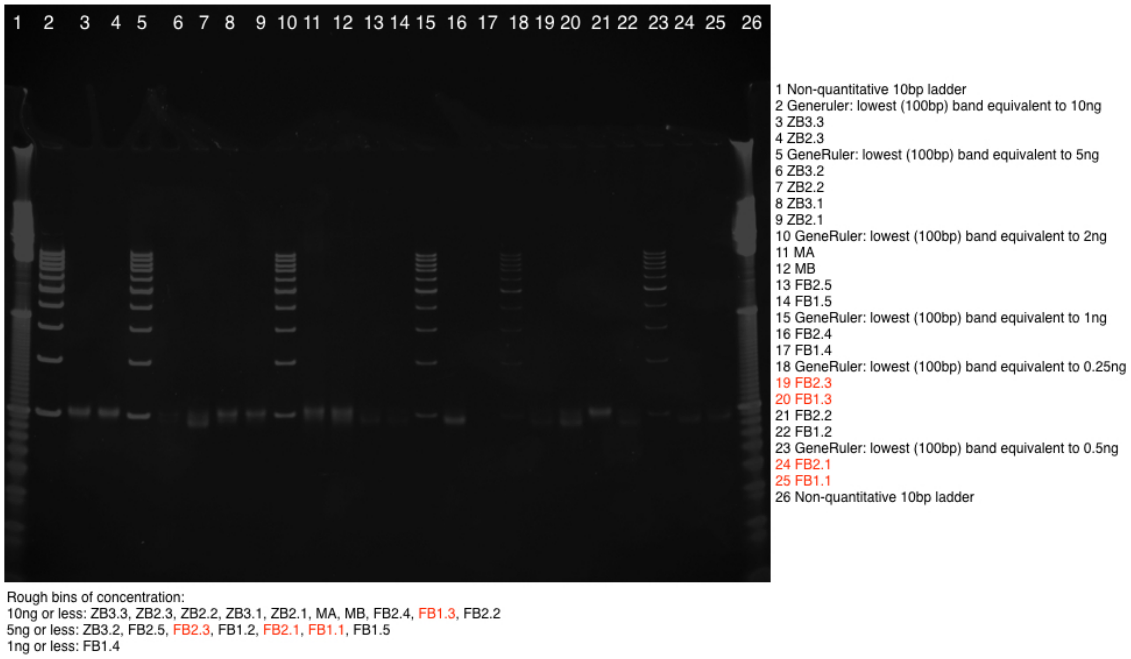


Figure 3.5. **Libraries and approximate concentrations.**

In order to best approximate the concentration of each of the cDNA libraries, aliquots were electrophoresed alongside known concentrations of GeneRuler ladder. The libraries' concentrations were thus then estimated visually and placed into bins of 0-1ng, 1-5ng, 5-10ng. These assignments were blind-tested by two additional colleagues (Ajioka and Micklem, personal communication). Lanes and samples denoted in red are the ones that were taken forward for further experiments. FB1.1 and FB1.2 denote replicates uninfected cells; FB 1.3 and FB2.3 denote replicates of infected cells.

These cDNA libraries were then sequenced, on the Illumina GA platform at the University of Tokyo, by the late Dr Junichi Watanabe. FB1.1 and FB2.3 were sequenced on the same flow cell, while FB2.1 and FB1.3 were sequenced together on a separate one.

3.3 Results

3.3.1 Quality

Illumina GA sequencing of the small RNA libraries yielded raw datasets of between eight to ten million reads each. Of these, a small proportion (<3 per cent) included bases whose quality was too low to be assigned (depicted as "." or "N", indicating an unknown nucleotide): these reads were discarded from further analysis. Since the datasets represented raw reads from the machine, the first pre-processing step was to determine the number of unique reads in each dataset. Interestingly, for each of the datasets, it appeared that the ~10 million reads in each dataset collapsed to a relatively few ~2.5 million.

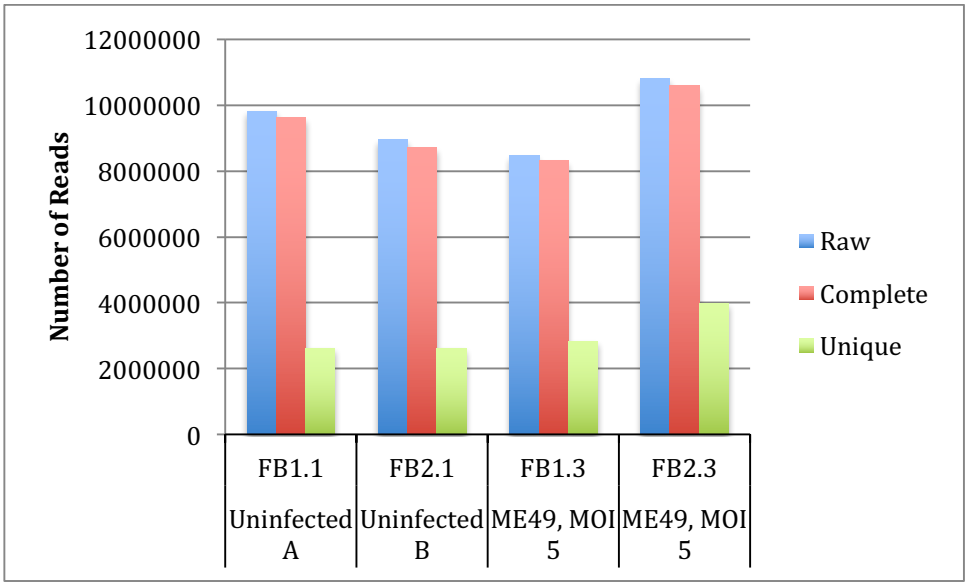


Figure 3.6. Chart of library sizes

Looking at the frequency distribution uncovered that in fact, the overwhelming majority of reads had been sequenced fewer than 1000 times. In

all four libraries, it was found that over 99.9 per cent of the reads had been sequenced fewer than 1000 times each. Further examination of the datasets (Figures 3.7-3.10) revealed that much of the apparent homogeneity in the dataset could be accounted for by a very small number of unique sequences: The seven most-frequently sequenced reads in each library accounted for between ~1.5 and nearly 3 million reads in each dataset, corresponding to between 14 and 31 percent of each library. This observation led me to examine these most-frequently sequenced reads manually, to see if I could determine their origin. In the most extreme dataset (FB1.1), all seven of these reads could be traced to either adaptor-dimerisation or ligation of an adaptor to a polyadenylation signal (Figure 3.7). In the other uninfected control sample, the situation was less clear-cut, with adaptor-dimers, polyadenylation making an appearance but also the highly-expressed mouse microRNA let-7. Let-7 also appeared in the datasets from infected samples, but two other sequences (with no significant alignments to the mouse genome but significant alignments to the *T. gondii* genome) also featured prominently.

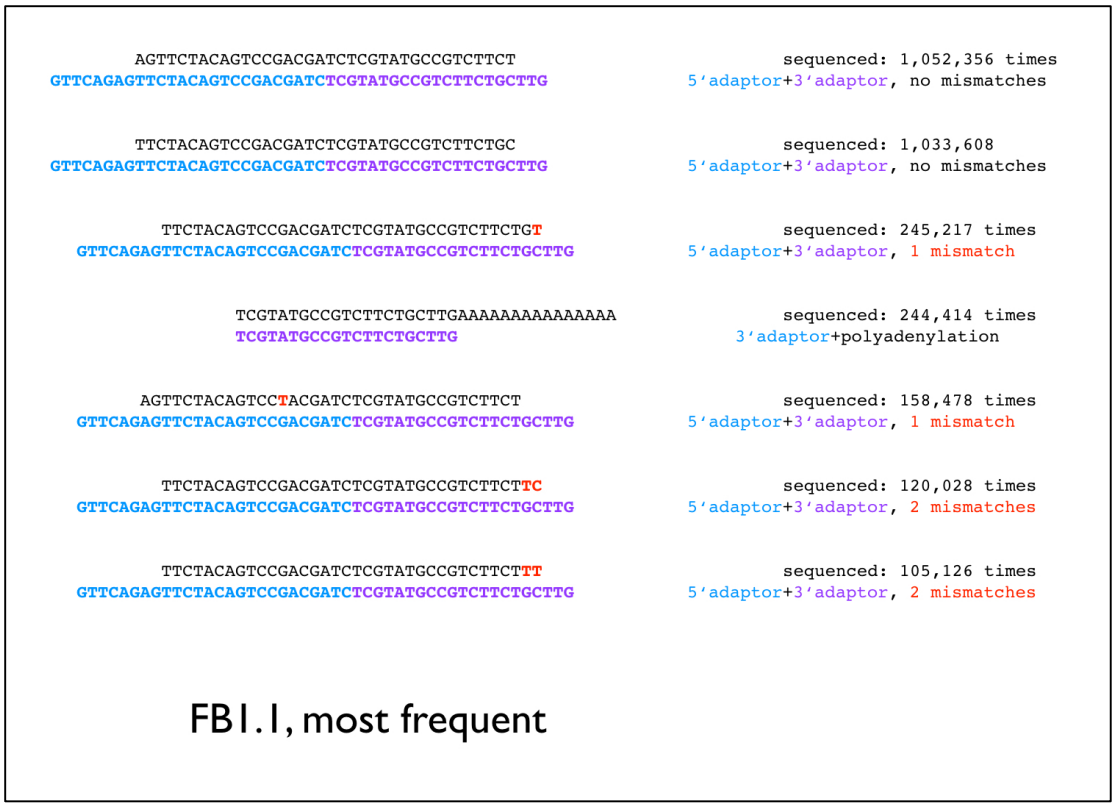


Figure 3.7. The most frequently sequenced reads in FB1.1 show significant adaptor contamination.

The seven most-commonly-sequenced reads in the first replicate of the uninfected sample were examined manually, to ascertain what their composition might be. In all cases, the sequenced reads could be accounted for through some combination of adaptor concatamerisation or ligation of an adaptor to a polyadenlyation signal.

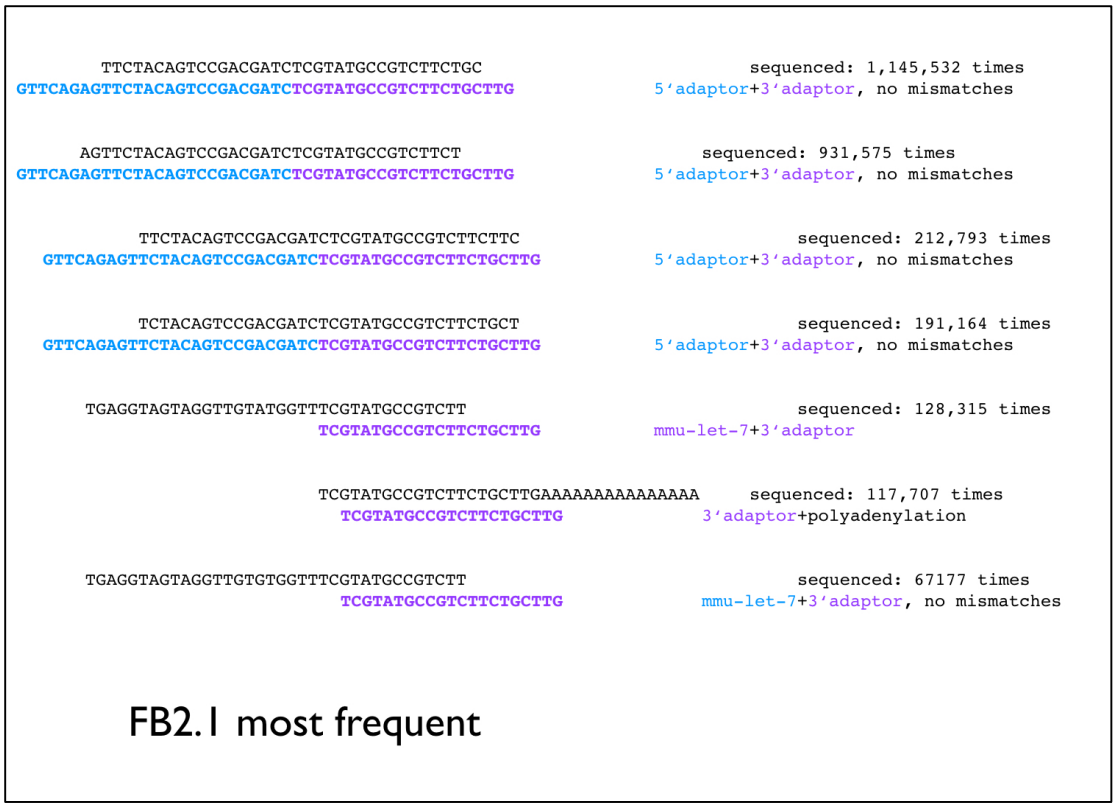


Figure 3.8. The most frequently sequenced reads in FB2.1 reveal adaptor contamination as well as let-7 miRNA.

The seven most-commonly-sequenced reads in the second replicate of the uninfected sample were examined manually, to ascertain what their composition might be. Unlike the previous replicate, there were two cases in which a highly expressed miRNA was present. In all other cases, as in FB1.1 (Figure 3.7), the sequenced reads could be accounted for through some combination of adaptor concatamerisation or ligation of an adaptor to a polyadenlyation signal.

Figure 3.9. The most frequently sequenced reads in FB1.3 reveal adaptor contamination, let-7 miRNA and unkonwn sequences.

The seven most-commonly-sequenced reads in the first replicate of the infected sample were examined manually, to ascertain what their composition might be. Two cases were not found upon a search of BLASTN. Apart from that, the most frequently-sequenced reads were accounted for by adaptor concatamerisations, adaptor-polyA ligations or known miRNAs.

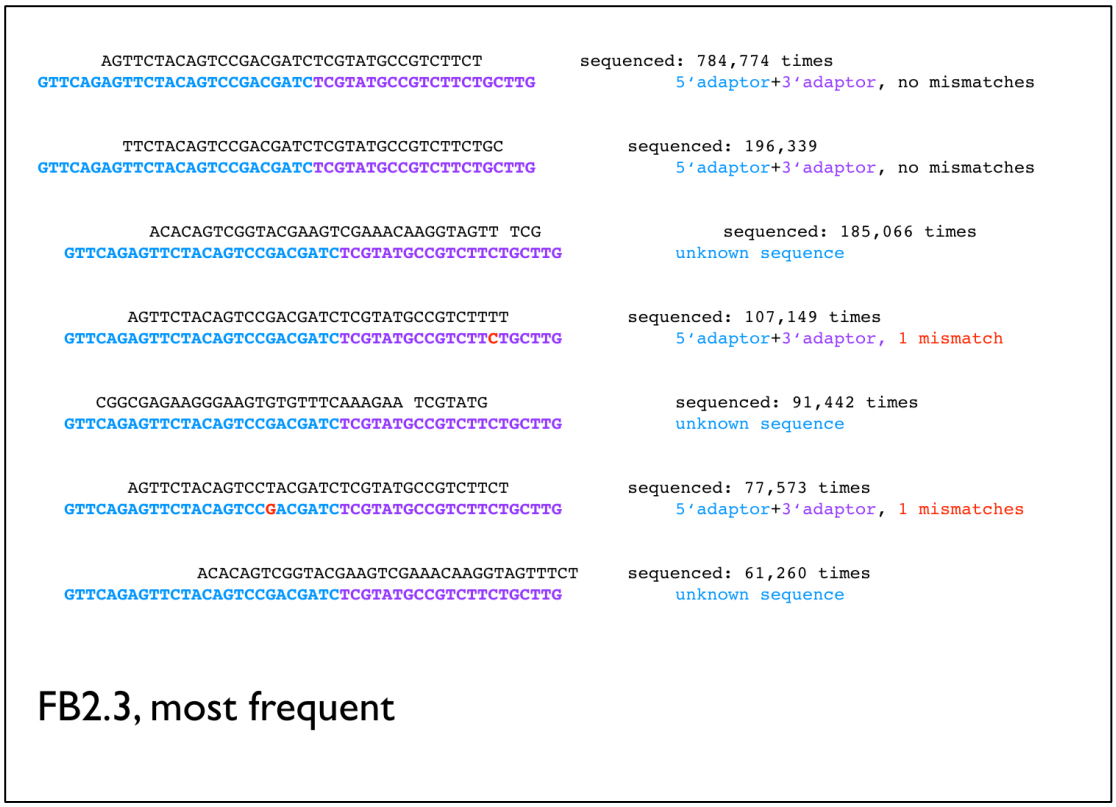


Figure 3.10. The most frequently sequenced reads in FB2.3 reveal adaptor contamination and unknown sequences.

The seven most-commonly-sequenced reads in the first replicate of the infected sample were examined manually, to ascertain what their composition might be. Two cases were not found upon a search of BLASTN. Apart from that, the most frequently-sequenced reads were accounted for by adaptor concatamerisations, adaptor-polyA ligations or known miRNAs.

Error

The fact that adaptor-dimers, with some mismatches, made up a large portion of my datasets – and that the sequence of these adaptors is known *a priori* – led me to develop a method to assess the error rates of the sequencing process. This error rate would then inform the parameters chosen for alignment of the reads to the mouse genome.

I created a set of short seed sequences by sliding a 9-mer window along the full adaptor-dimer sequence (blue and purple in Figures 3.7-3.10) and checked every resultant sequence tag against the full set of known mouse stem-loop pre-miRNAs (in either orientation on either strand) as well as

against each of my datasets. The most frequently-seen (in the datasets) 9-mer that did not match any part of any known mouse miRNAs was TCTACAGTC (a portion of the 5' adaptor). I used this sequence as a signature tag to identify adaptor dimers, and performed an exact search for it in my datasets, the results of which are shown in Table 3.1.

Table 3.1. High prevalence of the 5' adaptor 'tag' in each library.

Given the obvious contamination of my sequencing libraries by adaptor dimers containing the 5' adaptor, I performed a search for a 9-mer containing adaptor sequence within each. In terms of raw reads, close to half of each library contained this sequence, though this figure ranged from 7.6 to 12.5% when looked at as unique (collapsed) reads.

Presence of TCTACAGTC tag	Total reads	Unique reads
	Percentage	Percentage
FB1.1	5,189,994	327653
	54.00%	12.50%
FB2.1	4,132,876	303,535
	47.50%	11.60%
FB1.3	3,347,642	275,351
	40.20%	9.70%
FB2.3	3,396,108	304433
	32.10%	7.60%

I then examined the 'tails' that followed these nine bases. Overwhelmingly, the consensus sequence among the tails consisted of, as expected, bases corresponding to the remainder of the 5' adaptor and then the 3' adaptor but there was nevertheless quite a lot of variability, and this variability is what I chose to exploit in determining an error rate.

For each of the reads that contained the TCTACAGTC tag, I performed alignments to the reference (adaptor-dimer) sequence and recorded the number of mismatches for each read. The collated per-read or per-tail

mismatch rates for each dataset are shown in figure 3.11. The error rates appear more consistent between samples sequenced together than between biological replicates (FB1.1 and FB2.3 vs FB2.1 and FB1.3).

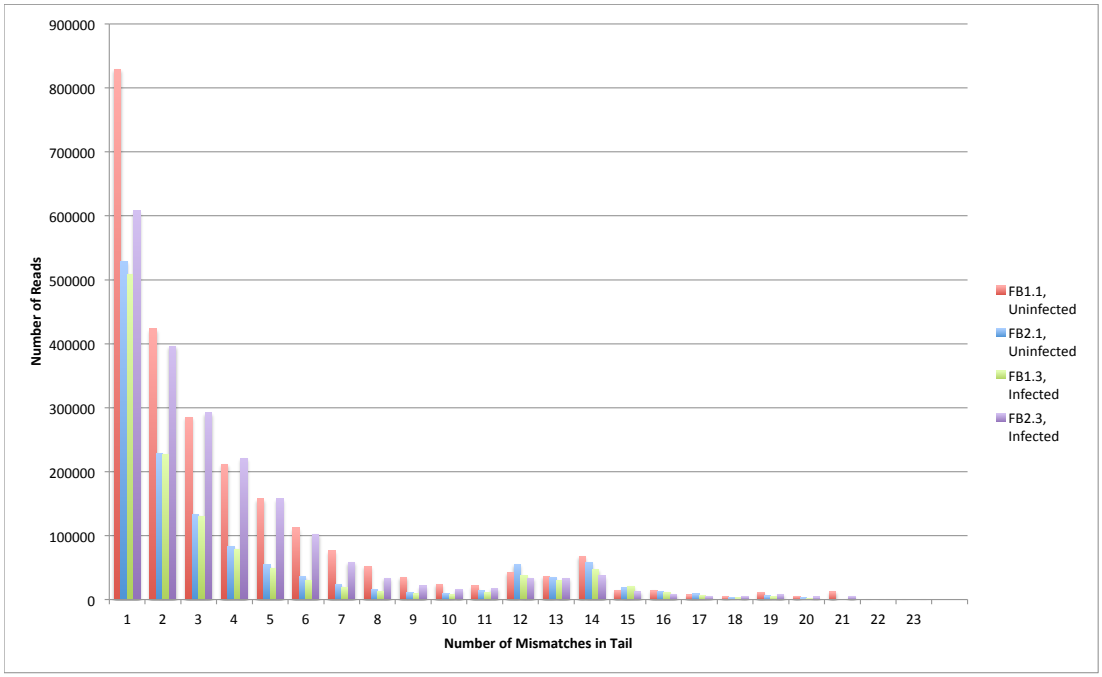


Figure 3.11. Mismatch analysis between the actual and expected tails following the 5' adaptor tag reveals sequencing batch effects and an error rate of >2nt.

Following the identification of the 9-mer 5'-adaptor tag (Figure 3.10), the tails following this tag were examined in each of the seqenced libraries and mismatches tallied up. This revealed an average error of 1.78nt.

In three of the four datasets, the average number of mismatches per read was less than two. The average across all samples was 1.78 and as such, I used a mismatch allowance of two in subsequnt alignments. [While in FB2.3, the average number of mismatches was 2.34, using a higher mismatch allowance of three for its subsequent genome alignments did not prove to be much of an advantage in terms of genome coverage, but only served to vastly increase the amount of time taken for both the alignment and subsequent analyses].

3.3.2 Removal of Adaptors:

Since miRNAs are shorter than the full length of an Illumina sequence read (36nt)[3], the issue of adaptor-removal from sequence reads is two-fold. First, a trace of the 3' adaptor should be a requirement for a read to be deemed as having come from a miRNA (rather than, for example, a degraded mRNA or a pre- or pri-miRNA). Second, these traces of adaptor must be removed before reads can be aligned to the reference genome.

The most common method of adaptor removal (those suggested in the MirDeep2 package (104) or the UEA sRNA ToolKit (105) for instance) consist of simple, exact searches for a specified minimum number of nucleotides matching the adaptor, and trimming reads accordingly. These methods, though simple to implement, have a number of disadvantages.

The quality of reads produced through deep sequencing – especially from earlier generation machines – tends to drop towards the 3' end of reads (117) and, since this is where the adaptor is expected to be located, searches for exact matches may miss a significant number of adaptor fragments. Moreover, given that the presence of a 3' adaptor fragment should be a requirement for inclusion of a read in downstream analysis, by missing adaptor sequences, simple searches may disqualify reads that in fact do contain 3' adaptor but have been mis-sequenced to a certain extent.

For sequencing runs where the quality remains high even towards the 3' end of the read, or for very large datasets that can afford such losses, these types of approaches might be satisfactory. For datasets such as mine, however, from earlier generation machines with un-optimized runs, allowances for sequencing errors are needed. The risk of allowing mismatches in adaptor-retrieval is that a 'nonsense' sequence or a degradation product might be incorrectly identified as adaptor and thus removed, with the remainder of the read being retained for downstream analysis. However, this can to some extent

[3] The longest mouse miRNA in mirBase version 19 is 27nt

be accounted for: if indeed a 'nonsense' sequence fragment was falsely identified as adaptor and trimmed, it is likely that the earlier portion of the read (the supposed microRNA) would also be nonsense. If so, it would be unlikely to align to the reference genome. If it did, or if the fragment originated from a degraded mRNA, it would not align according to the specified pattern expected of miRNAs. In subsequent alignment and coverage-determination steps, therefore, it would eventually be discarded anyway. The alternative – to discard reads from the outset because they do not contain an exact match to the 3' adaptor – is to exclude them from 'redemption' through alignment at all.

Another drawback of simple searches is that they fail to detect instances of adaptor concatenation, unless explicitly described and searched for. And, as shown in (Figures 3.7-3.10), due to adaptor truncation and the presence of some mismatches, many variations of adaptor contamination exist: identifying and specifying each of them in turn to then search and trim is laborious and sure to miss some instances.

One program that addresses both the issue of mismatches and of adaptor dimerisation is ScreenLinker[4]. This program aligns adaptor sequences to sequence reads using an "end-gaps-free" alignment, which aligns the entirety of one sequence (the adaptor) against another (the read under consideration), but does not penalise mismatches at either end. This type of alignment, while foremost searching for the 'expected' situation of an overlap between the beginning of the 3' adaptor sequence and the end of the read, also can be made to look for overlaps between the 5' and 3' adaptors, as well as concatamers thereof. Effectively, each read is screened twice, for each adaptor sequence provided. In the aligned region, a certain number of mismatches and indels are also allowable. However, screenLinker's scoring, while it allows for mismatches, also allows for indels. These are known to be far less frequent in

[4] The ScreenLinker program was a kind gift from Gordon Brown, CRUK.

Illumina datasets than those from other sequencing platforms and might render the programme too stringent, categorising sequences as being entirely made up of adaptor when they are not.

Thus, there exists a delicate balance between an allowance for mismatches and stringency when it comes to excluding adaptor sequences. Indeed, whichever adaptor-trimming method is used, the 'test' is the alignment pattern that it produces.

The strategy that I ultimately opted to employ to search for and remove adaptor fragments was a combination of the above methods. First, I aligned the raw reads against the adaptor sequence, using an exhaustive 'affine:overlap' model[5], which requires the inclusion of the start or end of the query (raw read) and the start or end of the target (adaptor sequence). This model of alignment gives rise to successful alignments following two possible configurations: Either the alignment would place the adaptor at the 3' end of the read (the desired situation) or at the 5' end of the read (spurious match or 5' adaptor concatemerisation). To distinguish between these two cases, I filtered the successful alignments on both the position of the adaptor match and the length of the sequence preceding the match (the length of the presumed miRNA). This enabled me to retain only those reads that followed the pattern: read—3' adaptor, with a minimum read length of 16nt.

Following this, I screened the resultant reads for 5' adaptor contamination, by applying a 12nt window along the length of the 5'adaptor and searching for these sequences within the clipped set. Reads that produced alignments to any 5' Adaptor 12-mer were discarded.

One interesting though little-discussed problem with adaptor removal is the issue of whether to retain or discard the final nucleotide when a read ends in T (the first nucleotide of the 3' adaptor). Arguably, this is more of a concern when it comes to first-generation sequencing runs where reads were 36

[5] This alignment was performed using exonerate (282)

nucleotides in length but, with the longest miRNA in miRBase being reported as being 34 nucleotides long, it is nevertheless an issue worth considering, especially if datasets are to be used for novel miRNA discovery in species where the 'rules' have not yet been fully characterised. As with many aspects of next-generation sequence manipulation, this issue is perhaps best thought of in a cost-benefit manner: The risk of allowing an adaptor-derived trailing T to stand might mean the difference between the read aligning to the genome or being discarded. If, on the other hand, single trailing Ts are removed but derive from true miRNA sequence, the remaining read will still align to the correct locus, albeit with a slightly less-well-resolved 3' end. These alignments will not affect differential expression of known miRNAs (no currently-described mouse miRNAs are that long) nor will they affect novel miRNA prediction, where precursor sequences flanking the putative miRA are excised and subjected to folding and other tests anyway. For this reason, I retained reads ending in U (T), but clipped this final nucleotide.

I sought to assess my method of adaptor removal empirically, by comparing it to the two other ones, by applying each of them to one of my datasets (FB1.1).

For the Simple Search method, I used a naïve regular expression search (as suggested by Krys Kelly, personal communication), to extract reads that contained at least one nucleotide of sequence (A, T, C or G) followed by the first eight nucleotides of the 3' adaptor sequence while also requiring the presence of this pattern. Of course, this method retains all lengths of remaining sequences (including, for example those instances where the eight bases of adaptor occur a single base in from the 5' end of the read). Most aligners typically have some minimum number of bases that they can usefully align to a reference and furthermore, the shortest known mouse miRNA is 16nt long. Thus, I excluded from the dataset any trimmed reads that were shorter than this.

I fed the same initial dataset to the screenLinker programme, using its default mismatch/indel score cut-off of 0.85, again with a minimum retained length of 16nt.

The number of unique and total reads retained using each of these three methods is shown in Figure 3.12.

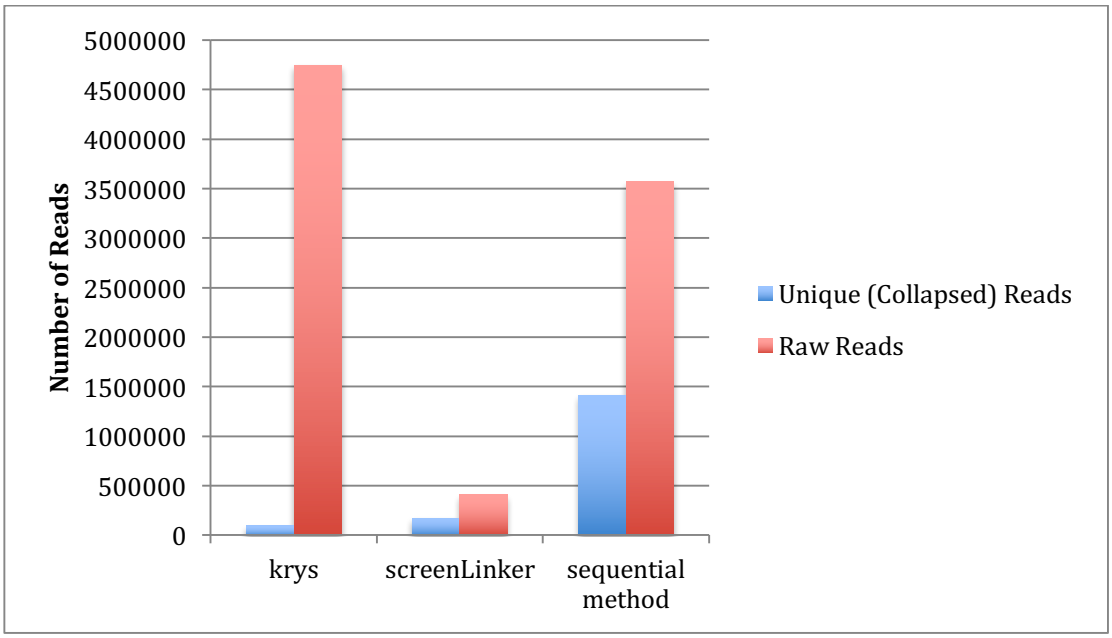


Figure 3.12. A sequential adaptor removal method yields a higher raw:unique read ratio compared to a naive regular expression search or screenLinker programme. After performing adaptor removal using one of the three methods described in the text (naive regular expression search, screenLinker or my sequential removal), read numbers were assessed. The sequential method yielded a moderate raw library size but a greater diversity of reads.

Prioritising only the number of retained reads is not a wholly satisfactory way to choose an adaptor removal strategy, however, as this criterion gives no indication of how the retained reads will ultimately align to the genome. Thus, I then used Bowtie (118) with a mismatch allowance of two to align each set to the mouse genome, the results of which appear in Table 3.2.

Table 3.2. Alignment statistics of pilot adaptor clipping – FB1.1 clipped according to one of the three methods

Reads clipped using one of the three adaptor-trimming methods were aligned to the mouse genome using Bowtie with a mismatch allowance of 2. The sequential and naive regular expression methods yielded high raw alignments but again, my sequential method yielded a greater diversity of aligned reads (more unique reads).

	Simple	ScreenLinker	My method
Aligned Reads - Unique	68,160	130,181	335,960
Aligned Reads – Total/Raw	3,186,788	340,871	1,649,877

As a control, I then filtered the successful alignments on their location, looking at matches to the locus encoding one of the most ubiquitous miRNAs: let-7a[6]. Using the alignment pattern to a well-characterised miRNA enables us to evaluate the pattern of alignment produced by each of these methods: whether they retain reads that align indiscriminately to the genome or whether they retain reads that have likely come from true miRNA loci (Figure 3.13).

[6] This was done using the coverage programs described in Chapter 4

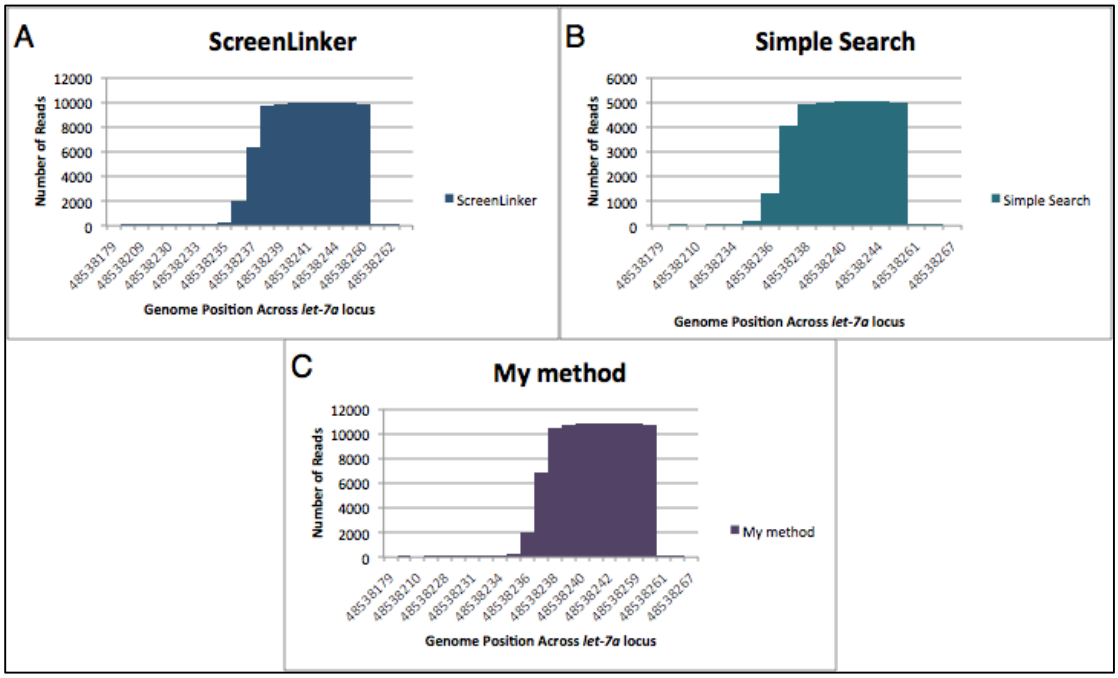


Figure 3.13. Coverage of the *let-7a* locus using different adaptor trimming methods reveals that a sequential method of adaptor removal performs best.

The well-characterised *let-7a* locus was probed after 3' and 5' adaptors were removed with either a pre-written programme ScreenLinker (a), my own sequential method (b) or a simple search (c), all described in main text. The ScreenLinker and sequential method performed similarly, with my own revealing slightly higher read counts being retained. NB the very small (but nevertheless present) proportion of coverage shows the putative *strand locus.

While my method and the screenLinker method appear to be fairly similar according to the metrics tested above, with screenLinker being slightly more stringent as expected, the simple search produces some puzzling results. It retained the most raw reads (Figure 3.12) and also produced the greatest total number of aligning raw reads, though these alignments came from the smallest number of unique sequences (Table 3.2). The alignment pattern along the let-7 locus provides more insight into how this might have occurred. Simple searches appear to retain a relatively small number of highly-sequenced reads. However, this type of adaptor-removal procedure completely disregards the issue of 5' adaptor contamination and this contamination, when allowed to proceed through alignment, can lead to spurious matches to the genome

(especially when alignment allows for mismatches). Indeed, the most highly-represented read retained by the simple search that also aligned to the mouse genome was the trimmed read "TTCTACAGTCCGACGATC", which the simple search had retained as though it were a true miRNA. However, this 18mer is in fact part of the sequence of the 5' adaptor however, and a clear result of adaptor concatamerisation[7]. This is further borne out by looking at the coverage profile of let-7a, where the simple search produced the fewest matching alignments: though more reads in total may have been retained, their pattern of alignment is clearly sub-optimal when it comes to identifying patterns of miRNA read provenance.

Having chosen an appropriately-balanced adaptor-removal strategy, I then applied this to the other datasets (Figure 3.14).

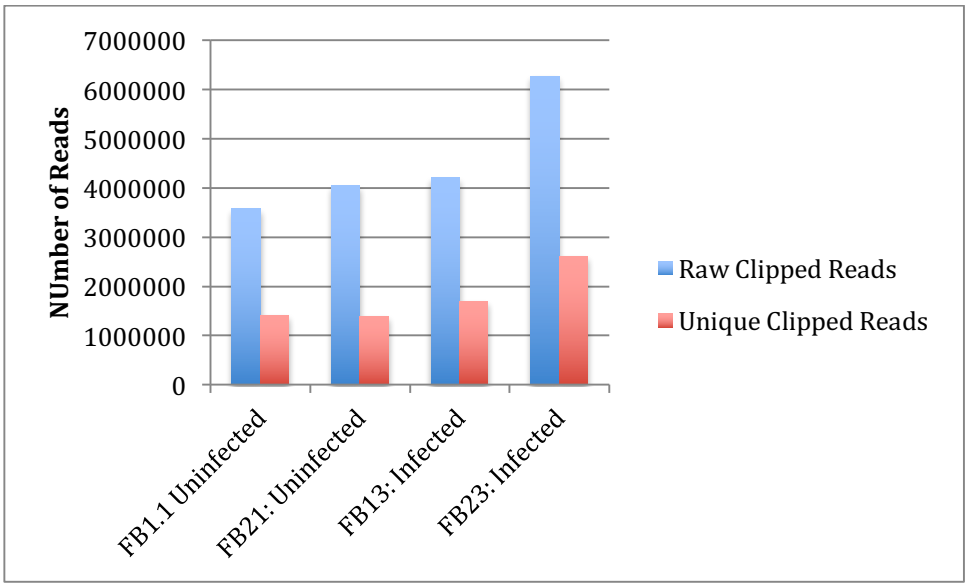


Figure 3.14. Number of reads retained from each library after adaptor clipping.

After using the sequential adaptor removal method described above, the number of raw and unique (collapsed) reads were tallied. All libraries showed a retention of over 1 milion unique reads.

[7] In fact, this is another indication that the method of relying only on alignment to screen out adaptor sequences is insufficient to exclude them.

3.3.3 Alignment

I used the per-read error calculations above to produce genome alignments for each dataset, using Bowtie (118). Alignment statistics for each dataset are shown in Figure 3.15. Rather few reads ended up aligning to the *M. musculus* genome – between eight and 19% of the unique trimmed reads. When looked at under the lens of raw reads (Figure 3.16), the situation is slightly better with most alignment percentages doubling. That being said, the number of unique mapping reads themselves is also quite small – under 300,000 in all libraries.

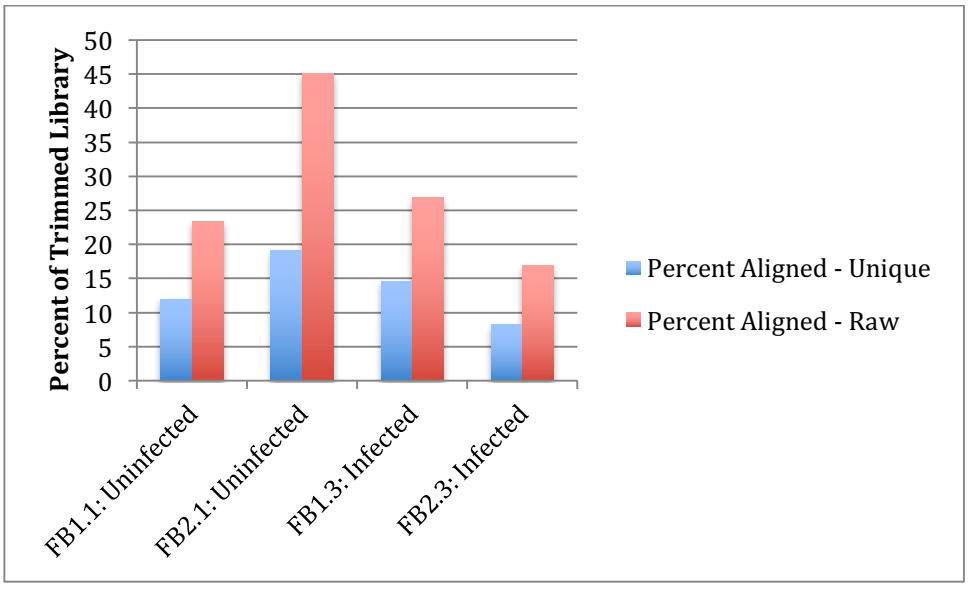


Figure 3.15. Reads that aligned to the genome, as a percentage of the trimmed population. The adaptor-removed reads were aligned to the mouse genome, using Bowtie (119). Between eight and 19% of the unique (collapsed) reads were retained after alignment, while between 23 and 45% of raw reads were retained.

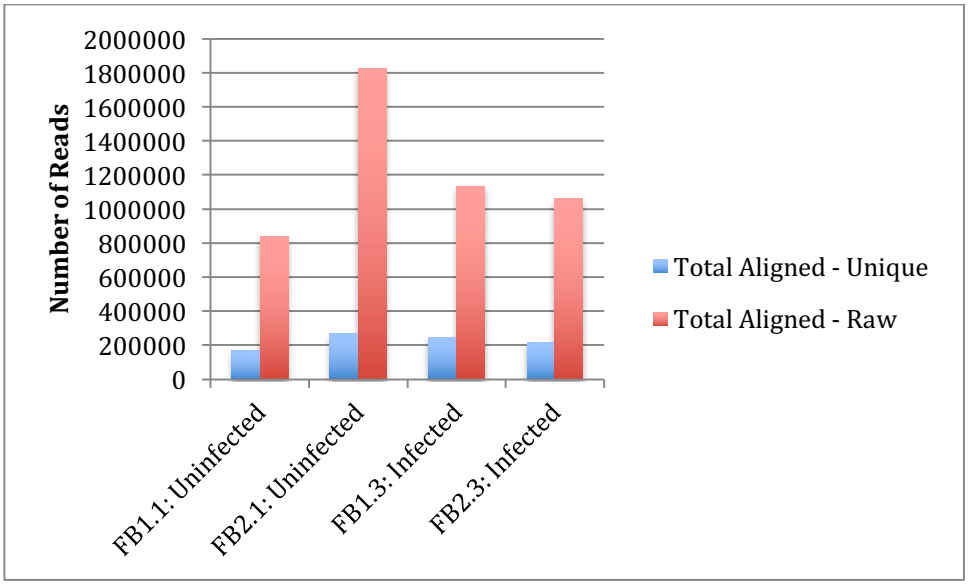


Figure 3.16: Raw number of reads that aligned to the mouse genome.

Adaptor-trimmed reads were aligned to the mouse genome using Bowtie (119). Aligned population sizes ranged from 170,762 to 268,739, whereas these same libraries accounted for as raw (uncollapsed) reads ranged from 838,860 to 1,827,933 reads per library.

Reads that aligned to the mouse genome according to these parameters were then selected either for expression analyses of known mouse miRNAs (**Chapter 4**) or for the identification of novel miRNAs.

3.4 Discussion

3.4.1 Novel microRNAs in Mouse only

In order to predict novel miRNAs expressed in my dataset, I used miRDeep2 (113) to analyse my aligned reads. At a score ≥ 0, (an estimated signal-to-noise ratio of 1.5 or above), 415 of the 1281 miRNAs in miRBase were picked up in my datasets by miRDeep2 (81%).

At this same score, 180 novel miRNAs were predicted over all four samples. Two had significant hits to rRNA or tRNA and were thus excluded from consideration as potential mouse miRNAs. 34 of the remaining putative miRNAs were scored using read-evidence from the infected samples only, which could be explained in two ways: either these are *T. gondii* miRNAs which also map to the mouse genome or, these could be mouse miRNAs

expressed only under conditions of infection. Given that potential novel miRNAs are scored not just on sequenced reads but also with a strong emphasis on genomic context (and the folding of that context), it is unlikely that these miRNAs derive from the parasite. Nevertheless, in order to be as sure as possible of their origin, I aligned them to the *T. gondii* genome. 26 of them were found to align and could thus not, on the basis of alignment alone, unambiguously be deemed mouse miRNAs. The remaining (which did not produce alignments to the parasite genome) were added to the list of putative novel mouse miRNAs, now numbering 152. These miRNAs are listed in the Appendix.

3.4.2 Novel microRNAs in Toxoplasma only

During the preparation of the RNA samples for sequencing, it is impossible to exclude *T. gondii* RNA from that of the infected host and so, of course, a proportion of the sequenced reads will have come from parasite material. I thus sought to mine these reads for putative parasite miRNAs. Since reads mapping to both the mouse and *T. gondii* genomes are difficult to resolve based on origin, I used a subtractive method as a starting point. I aligned to the parasite genome (with the mis-match allowances of two) only those reads that did not map exactly to the mouse genome.

Since none of the miRNAs predicted be any previous study (106, 109, 120–122) have as yet been accepted for inclusion into miRBase, far less information about their characteristics is available. As such, the number of miRNAs returned by miRDeep2 is quite small (this is also a result of the small proportion of sequenced RNA having come from parasite material to begin with) and accordingly, their confidence level is lower.

As a control, I performed this same analysis on the uninfected samples, and excluded any putative miRNAs identified by miRDeep2 that were present in both infected and uninfected sets.

Six microRNAs were thus identified as putative *T. gondii* microRNAs at a score > 0. These miRNAs are listed alongside the putative novel *Mus musculus* ones in the Appendix.

3.5 Discussion

3.5.1 Library Preparation

It is clear that sequencing data produced by such methods as Illumina require careful handling and that there are many pitfalls. This is especially true of early-generation machines and processes, where many of the 'quirks' had still not been worked out. For instance, my datasets did not include any quality information (such as Phred scores), which would have aided greatly both in terms of adaptor trimming and in terms of alignment. The majority of alignment algorithms (at least in their more modern versions) allow for FASTQ scores to be taken into account.

Instead, I attempted to calculate an error rate based on the presence of adaptor-dimers in my libraries. In fact, there were a great many adaptor dimers – something that should occur far less with newer library preparation protocols. For instance, Illumina's TruSeq v2 Small RNA Prep Guide now includes an adenylation step to prevent concatamerisation. Other library preparation kits now include the use of hairpin adaptors (123), or the destruction of unligated adaptors at every stage (124).

3.5.2 Adaptor Removal

Even if adaptor dimers are correctly removed (or are not formed in the first place), the construction of sequencing reads still means that for small RNA sequencing, adaptor trimming is necessary – and even desirable, given the structure of the product to be sequenced. Nowadays, read lengths are longer than 36nt and so the confusion between the potential end of a miRNA and the beginning of the 3' adaptor is no longer such an issue. The issue of the

final trailing nucleotide being ambiguous between the miRNA and the beginning of the adaptor is also less of a risk. That being said, more sophisticated adaptor removal algorithms have also been developed: cutadapt (125) is one of these, that affords the ability to require the 3' adaptor and/or screen for 5' contamination as well as potential dimers.

3.5.3 Alignment

Since the emergence of NGS, several different alignment algorithms have been developed to aid in dealing with the large volumes of sequence read data. It could be argued however that the number of tools has grown at a faster pace than users' ability to evaluate them. And, a number of issues are still under-addressed. As Kang and Friedländer note, "Overall, the field of mapping sRNAs is understudied [...]" (126), a concern they raised particularly in the context of multi-mapping. There is some debate as to the number of times a read should be "allowed" to map to a reference genome, complicated by the fact that some miRNAs do in fact come from highly repeated genomic regions. There are several possibilities, which impact both the coverage of the genome and the interpretation of expression data from such alignments. Perhaps the most common (and arguably most stringent) method to deal with this issue is simply to exclude all reads that map to the genome more than a certain threshold number of times. Apart from having the obvious disadvantage of throwing away vast amounts of what might be useful data, this method is unsatisfactory in other ways. Most published papers that do this, for example, do not justify the selection of a particular cut-off, beyond a vague mention of reads aligning to rRNA regions. Moreover, these cut-offs vary widely, with some groups using a cut-off of one (only allowing reads that map to a single location) (127, 128), or five miRDEEP2 (129), and others extending this number up to 500 (130).

Some alignment programs attempt to address these concerns by providing options for selecting or reporting multiply-mapping alignments. SOAP (Short Oligonucleotide Alignment Program) (131), reports the "best" of all possible alignments for a certain read (i.e. the alignment with the fewest mismatches, or highest quality score) and, if there are more than one such alignments, the user can choose whether to include all, none or one picked at random. Similarly, Bowtie also gives the user these three options but does so for all valid alignments, not just the "best". So, if a read has two alignments with no mismatches and one with two mismatches (and a threshold has been specified such that two mismatches are 'allowed'), the user can then opt to report all three alignments, just the two best or a single one chosen at random (from either the two best or all three). That being said, the default behaviour of these algorithms is often unclear (random assigning is the default for Bowtie, for instance, although even this is not thought to be truly random but rather "pseudorandom" (132)). Of course there is no guarantee that any such alignment will be the correct one, and, even choosing to report the "best" alignments may not be adequate, especially if one considers the possible issue of miRNA editing. Two groups have sought to clarify this issue using local-weighting and probabilistic distribution of multi-mapping reads (133, 134), but despite their utility, it is likely that unless they are incorporated into existing miRNA analysis pipelines or easily-compatible standalone packages they will find a limited audience.

For my purposes, all alignments produced for the prediction of novel miRNAs would later be scored according to other criteria, such as similarity to known miRNAs and stringent folding parameters, and thus, I opted to accept all possible alignments, which is considered by some to be a more prudent choice, in the absence of better easy-to-use statistical methods (135).

The libraries I used had a large amount of adapter-dimerisation that needed to be filtered out, and of the remaining reads, only a fraction aligned

correctly to the genome. As a result, the novel miRNAs identified here must be treated with caution. The huge data generation possibilities afforded by NGS mean that the number of putative novel miRNAs in the sample space has exploded and miRBase, the foremost database for miRNAs has, through its self-admittedly "inclusive" (80) approach, likewise seen its population increase very rapidly. A tempting parameter in 'scoring' putative novel miRNAs is to use 'real' ones as a baseline (i.e. ones from miRBase). But, this automatically biases the criteria for inclusion. For instance, to take the multi-mapping scenario above: if one aligns all mouse miRBase miRNAs to the mouse genome with no mismatches, then it does appear that the vast majority of them align once and once only (data not shown). A naïve researcher may then use this as a criterion for scoring a putative novel one. However, if one considers *how* those miRBase miRNAs were themselves found (did the researchers who found them have a unique-mapping criterion?) the arguments underpinning several parameters quickly becomes circular. It will therefore be essential (as is beginning to happen (80, 136)), as more and more RNAseq experiments identify ever-greater numbers of putative novel miRNAs, to ensure that the bioinformatics approaches remain firmly rooted in the mechanisms of miRNA biogenesis and proceed alongside a well-curated reference set.

IV. Modulation of Host microRNAs

4.1 Introduction

4.1.1 MicroRNAs and *Toxoplasma gondii*

Despite the considerable interest that miRNAs have garnered in recent years, relatively few studies have looked directly at potential changes in the miRNA landscape of host cells following *T. gondii*.

One of the earliest was Zeiner et al's study (137), where the authors used miRNA microarrays to assess differences in miRNA expression at 6-, 12- and 24-hours after infection, as compared to changes at these time points in mock-infected cells. The experiment was performed in HFF cells, using RH at an MOI of 5. The limitations of microarrays in looking at miRNA expression are discussed in **Chapter 1**, but nevertheless, several miRNAs appeared to be dysregulated as a result of infection. The authors concentrated on a single polycistronic cluster of miRNAs (the miR-17~92 cluster) which was upregulated two-to-threefold in infected cells. This cluster encodes miRNAs whose dysregulation has been shown to have potent effects in oncogenesis; they are overexpressed in a vast array of cancer types. Its targets include such key regulators as *Pten* and *Bcl2l11*, and knock-out of this family's miRNAs has resulted in a pro-apoptotic phenotype. However, the regulation of and by this family is likely to be quite complex. The earliest association was with MYC, where the transcription factor has been shown to directly upregulate this miRNA family. Thus, the overexpression of miR-17~92 family members may be a result of parasite-induced upregulation of MYC (21). This may not be the only regulatory mechanism relevant in *T. gondii*-mediated modulation, however. This miRNA family is strongly repressed under hypoxia in a direct *Trp53*-dependent manner, with a concomitant strong sensitization to hypoxia-driven apoptosis. *Toxoplasma gondii* infection follows this pattern, given that it is known to induce and stabilise the expression of HIF1A even under

normoxic conditions (138), and also results in downregulation of TRP53 protein levels (19). It may therefore be that hypoxia-related signalling pathways are stimulated by the parasite but the concurrent downregulation of *Trp53* (see **Chapter 6**) places the upregulation of miR-17~92 solely under the control of MYC. In any case, it is clear that when unpicking the regulatory networks affected by *T. gondii*, miRNAs have a role to play.

To my knowledge only two other studies have looked at the impact of parasite infection on host miRNAs. The first is another global study, where Cai et al infected human macrophages with an atypical *T. gondii* strain designated China 1 and performed miRNA microarrays to asses the difference (139). Computational analysis of the upregulated miRNAs' pri-miRNAs showed that many of them (including the gene that encodes members of the miR-17~92 family) contained potential STAT3 binding sites. Moreover, siRNA against STAT3 blocked the expression of this pri-miRNA following infection. While it is unclear what the behaviour of STAT3 is in host cells infected by the China 1 strain (the authors' Western Blot only STAT3 protein levels dropping with siRNA transfection in the absence of infection), it is likely that STAT3's possible involvement adds yet another level of complexity to the regulation of this important miRNA family.

The final study dealing with host miRNAs in the context of *T. gondii* looked in great detail at two particular miRNAs: miR-146a and miR-155. They found that miR-146a was upregulated in Type II infection (expression was unchanged following RH infection, even though it was applied at the same MOI of 3 which implies a much higher parasite burden per cell at each of the time points). The researchers also looked at miR-155, and found that it was upregulated in both strains. Following these results, Canella et al then looked at miRNA expression in the brains of chronically-infected mice and found that both of these miRNAs were upregulated (as compared to uninfected mice). They were able to correlate high cyst burden in these mice

(differential cyst burden was induced by infection with different progeny of a IxII cross) with high levels of miR-146a. Moreover, mice in which miR-146a had been knocked out exhibited a greater resistance to infectious challenge with the PruA7 strain (Type II). This set of striking experiments make it clear that the modulation of host miRNAs by *T. gondii* can have far reaching effects *in vivo* – even beyond the life stage where differences were first assessed.

4.2 Methodology

I began my analysis with the aligned libraries from **Chapter 3**.

4.2.1 Coverage

Most alignment programs, however their mode of action, report successful alignments as start and end co-ordinates, both of the query sequence and of the match. For this positional information to be used either to identify putative novel miRNAs or to track expression of known loci, it needs to be translated into depth-of-coverage. That is, for a given window of interest along the reference sequence (usually the reference genome), we need to know the number of reads from the sequencing dataset that have mapped.

A naïve approach to this would be to create an array for each chromosome, consisting of as many elements as the chromosome has bases, with each array-element initialised to zero. Then, for each aligned read, for every position from the start to end co-ordinates of the match, the chromosome array would be incremented by one (Figure 4.1).

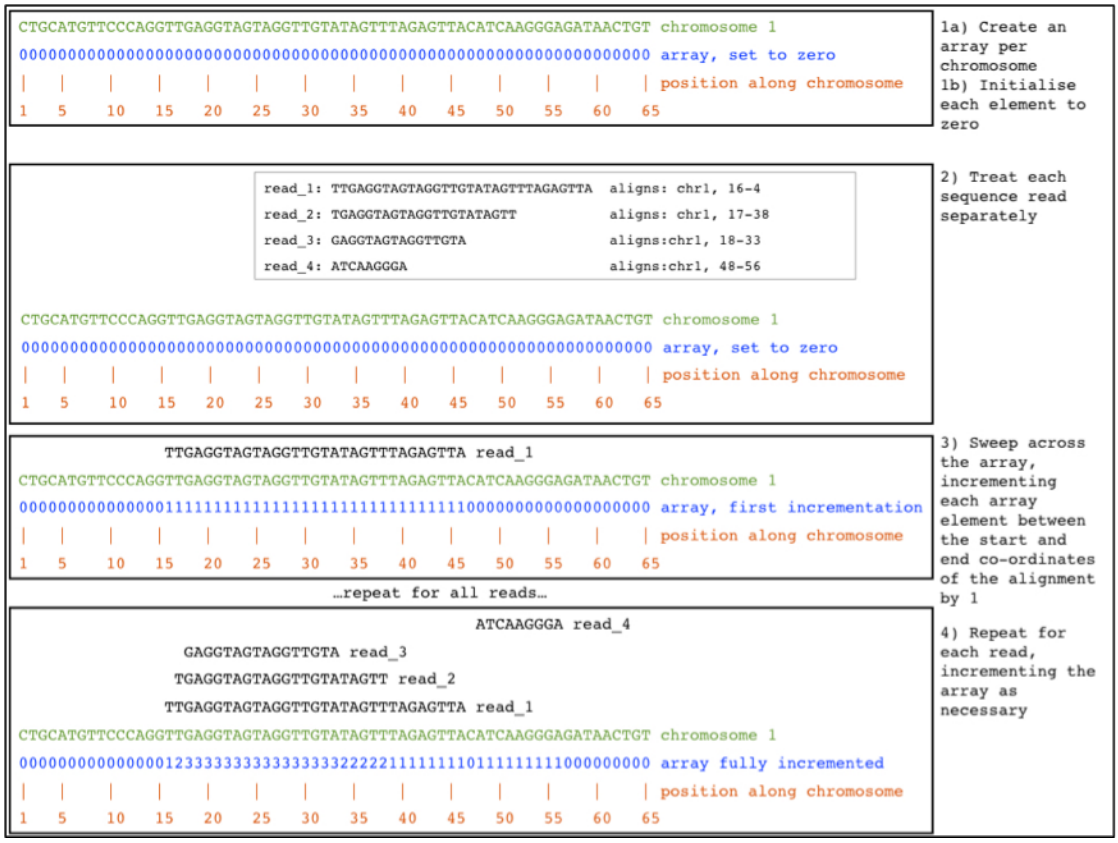


Figure 4.1. Generating coverage from aligned reads using a naïve method of array sweeping is computationally costly.

Using a 'dummy' set of aligned reads, generating coverage across a genome could be achieved by sweeping across a genome array and incrementing each position of each read – this is a computationally costly method, requiring 79 increment operations.

Effectively, this means that every nucleotide of every read needs to be treated separately, to appropriately increment the chromosome array. For large datasets, this approach can be very computationally expensive.

A faster method, with lower memory usage, implements the strategy of run-length encoding. Here, instead of recording and marking every base in each read, only the start and end co-ordinates of alignments are considered, as potential *positions of change* in coverage. This approach is likely to be effective for datasets – such as small RNA libraries – that are expected to align to a reference genome in a relatively sparse fashion, as oases of stacked reads in a desert of zero coverage. That being said, given the overlapping

nature of aligned sequencing reads, the framework for creating the run-length-encoded coverage array needs to be established with care, to avoid, for example, 'off-by-one' errors (OBOEs) or overlooking the starts and ends of the coverage arrays themselves. The steps that I took while constructing the program to create and populate run-length-encoded coverage arrays are shown in Figure 4.2.

For each aligned read, the start and 'end+1' co-ordinates were collected and sorted uniquely (such that, for instance, for several aligned reads with the same genomic start co-ordinate, that position would appear in the list only once: as a single *position of change*). To this list were appended the first position of the chromosome as well as a 'dummy' co-ordinate one position beyond the end of the chromosome (these two extra co-ordinates need to be defined explicitly since our reads are not guaranteed to overlap the chromosome ends). This list of co-ordinates was then used to create an empty array, with as many elements as there were changes in coverage. Each sequenced read is then taken in turn and the corresponding coverage elements incremented appropriately.

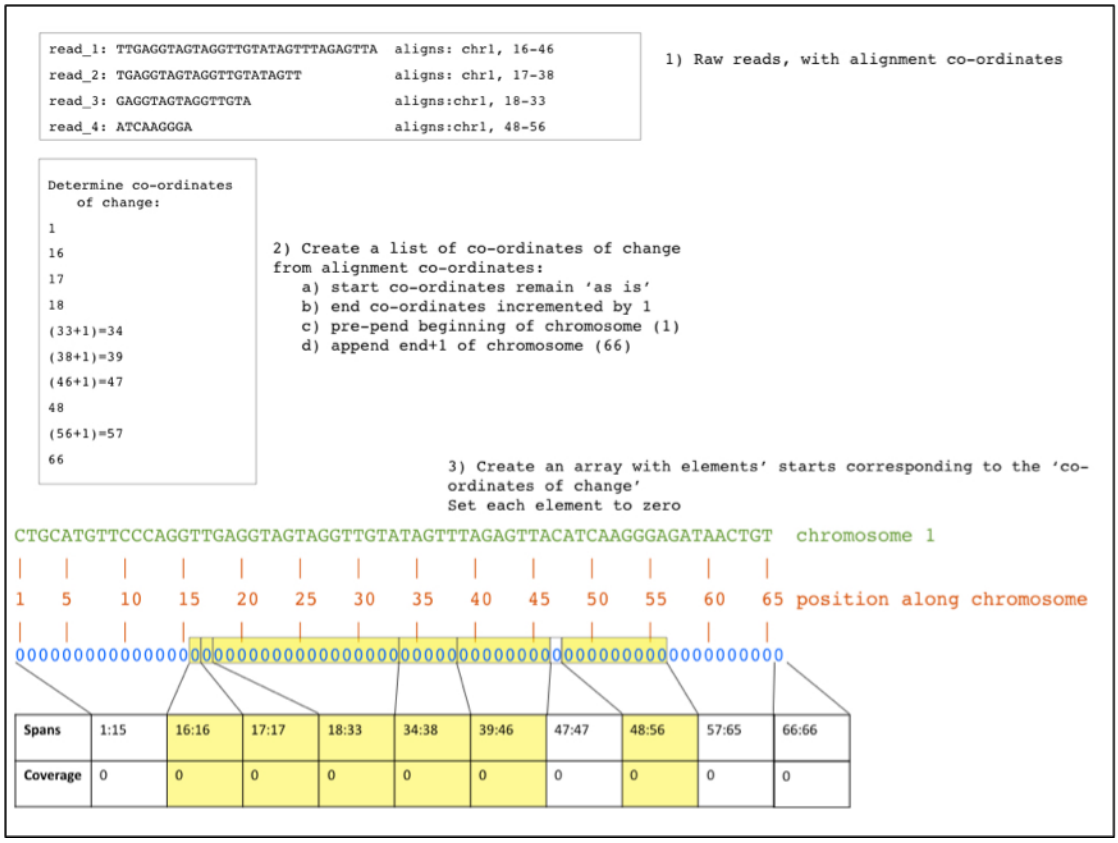


Figure 4.2 (a). Run-length encoding implementation to coverage generation

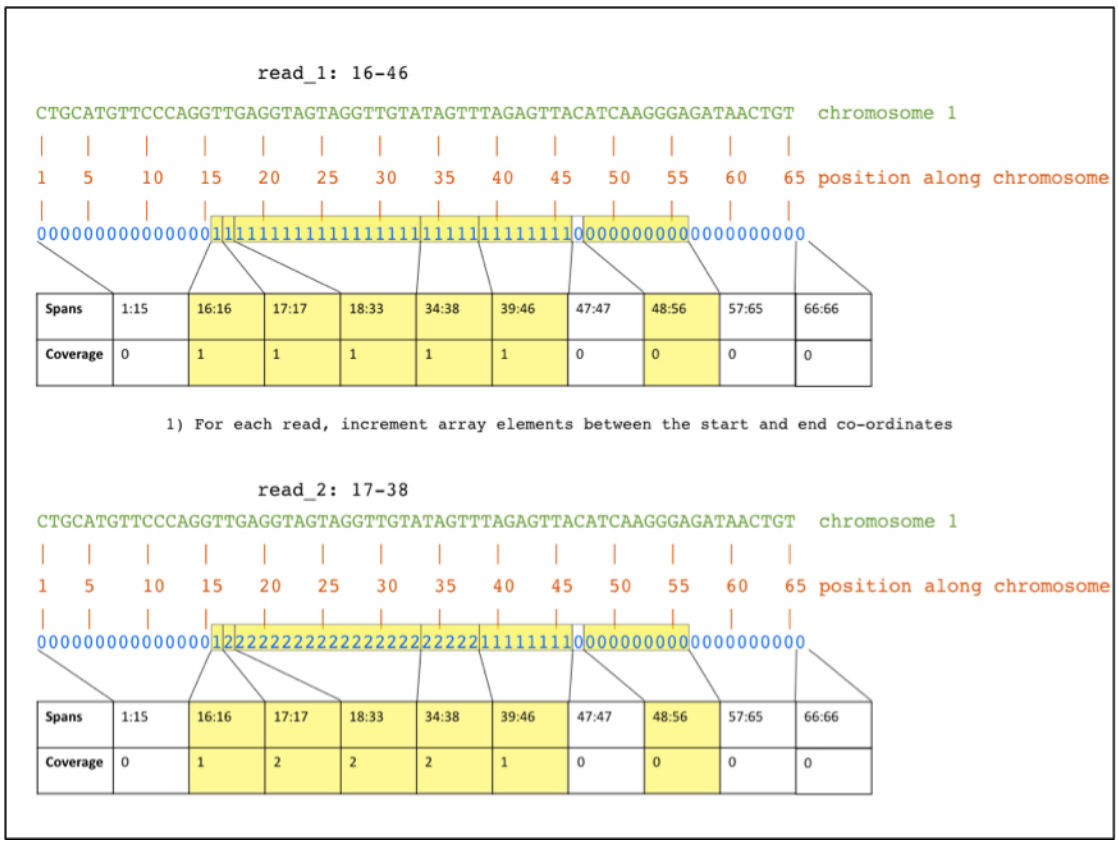


Figure 4.2 (b). Run-length encoding implementation to coverage generation

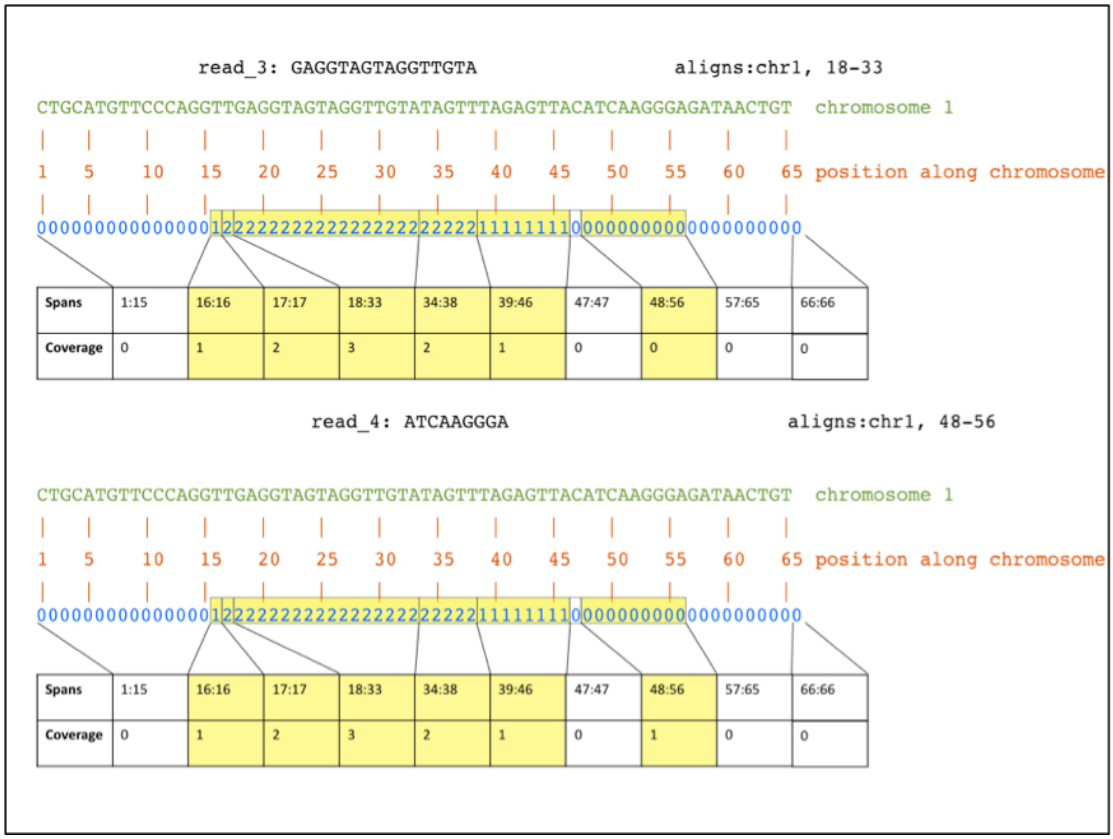


Figure 4.2(a-c). Implementing run-length encoding to generate coverage across a genome is more efficient than genome sweeping.

This schematic shows that by considering *positions of change* in read alignments instead of the entire length of each read, computational coverage generation can be made more efficient.

Using this method, for each chromosomal array, far fewer increment operations need be performed than with the per-base method. In the example in Figure 4.2, for instance, the per-base approach requires 79 increment operations (the entire length of each of the reads summed), scanning every position along the chromosome. Using run-length encoding on the other hand, this number is reduced to nine, only scanning and checking ten array elements per read.

4.2.2 Locating Features within Coverage Maps

Once coverage from alignments has been recorded, the next step is often to collect this coverage (the number of reads aligned) for specific regions of interest: known miRNAs, for example. Intuitively, once would begin at the

first position of the incremented coverage array and read across until the start co-ordinate of the region of interest had been reached and then record coverage information from there up to the end co-ordinate. Again this approach is time consuming and computationally expensive, as it requires scanning of the entirety of the array (chromosome) within which the target region lies.

A faster reporting method can be achieved by implementing binary searches. This approach splits the search field into halves and evaluates whether the target has been found (as being the central element), or whether this central position is too high (the target is to the 'left' of the central element) or too low (the target is to the 'right'). The half that is shown to contain the target is then itself split, and its central position, left and right halves are again evaluated. The entire process of splitting and evaluating is applied recursively until the midpoint of the sub-array and the target are found to be one and the same (Figure 4.3) This is applied to find both the start and end co-ordinates of the region of interest, which, once located, are used to look up the coverage information from the coverage array.

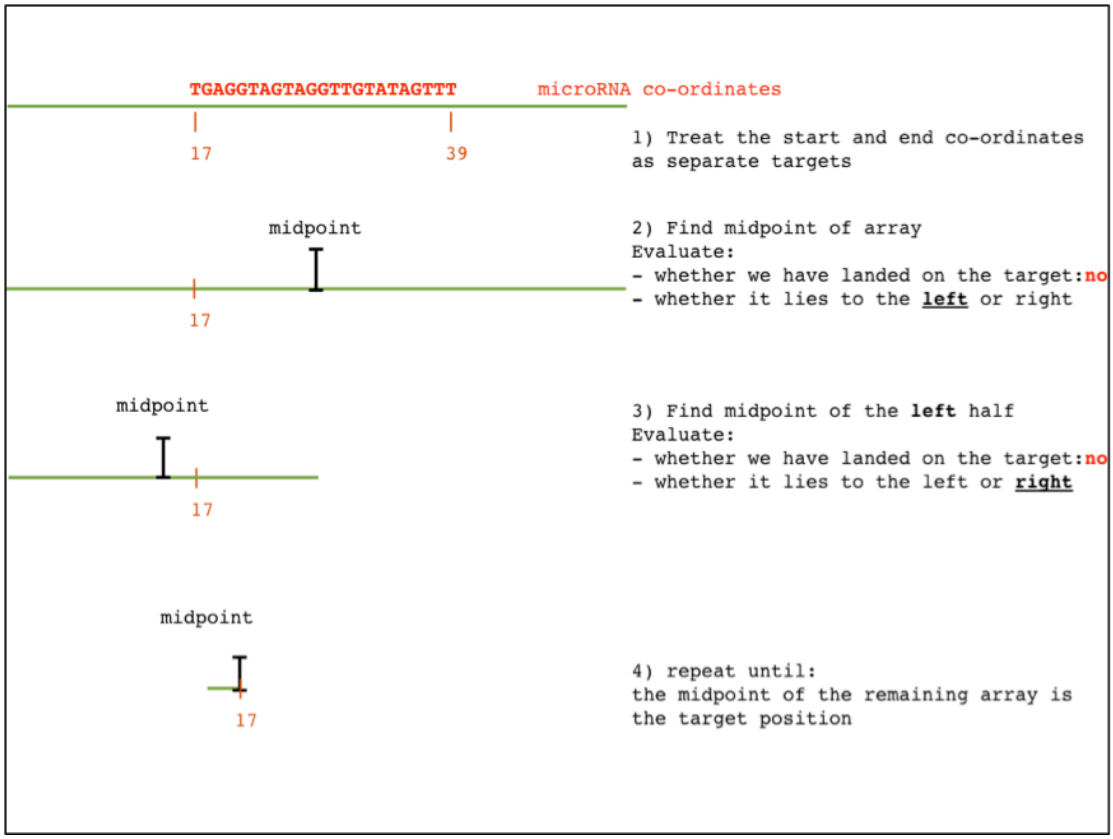


Figure 4.3. A binary search method is a rapid and efficient way to locate a featyre of interest within an array.

Once coverage maps have been generated, a binary tree method can be used to locate known end co-ordinates of microRNAs. In this method, each region of interest is split in two and each half is examined to see if the feature is present. The correct half is then itself split in half, recursively, until the feature of interest has been found.

This reporting method is thus used to generate tables of count data, to be then evaluated for the presence of differentially-expressed microRNAs.

4.3 Results

After aligning the sequenced reads to the genome (as in **Chapter 3** except allowing only a single mismatch per read), I then generated coverage for all known *Mus musculus* miRNAs according to their co-ordinates from miRBase, and produced tables of count data for known miRNAs. The highest scoring span across the entirety of the miRNA was taken as the count, for each miRNA.

4.3.1 Differential Expression Analysis

There exist a number of tools to assess the differential expression through deep sequencing experiments. While it was not designed for such small datasets such as mine, I nevertheless applied EdgeR (140). EdgeR uses a method of estimating dispersions in order to apply its statistical tests. These are difficult to estimate correctly when samples/replicates are highly divergent (as mine were). But, EdgeR allows for some batch effects, which I identified as being related to the flow cell each sample was sequenced on. So, I separated the replicates based on this factor to help correct for batch effects arising from this. Though 30 microRNAs were initially identified as being differentially expressed with a p-value <0.05, after multiple-testing correction at an FDR $<$ 0.05, this number fell to five (Tables 4.1 and 4.2).

Table 4.1. microRNAs upregulated following infection

miRNA	logFC
~~mmu-miR-3096b-3p*~~	~~4.9~~
~~mmu-miR-1935*~~	3.1
mmu-miR-200b-3p	5.9

* These two microRNAs have since been removed from the miRBase database as they were shown not to be genuine microRNAs.

Table 4.2. microRNAs downregulated following infection

miRNA	logFC
mmu-miR-3080-5p	-5.5

4.4 Discussion

There were several issues with this set of experiments, as was detailed in **Chapter 3**. Largely, the low quality and poor depth of coverage by the libraries had severe knock-on effects, leading to few confident assessments of differential expression. As a result, only an extremely limited picture of host miRNA regulation in response to *T. gondii* infection can be gleaned.

Of the four miRNAs found to be dysregulated following infection, two have subsequently been removed from miRBase (80). Mmu-miR-3096 family-members showed reads that were inconsistent with processing by the usual miRNA machinery, in a variety of recent validation experiments that used AGO-CLIP and Dicer knock-out to identify miRNAs whose expression is likely to have emerged from canonical processes of miRNA biogenesis (136). The other now-excluded miRNA, mmu-miR-1935 was found in the same study to be a repeated Alu/B1 SINE element, also transcribed in a method inconsistent with canonical miRNA biogenesis.

Thus, two miRNAs remained as being dysregulated by *T. gondii* infection, one downregulated and one upregulated.

mmu-miR-3080-5p

Unfortunately, mmu-miR-3080 is an extremely poorly characterised miRNA, not just in *Mus musculus* but in any species. It has been showed to be moderately upregulated in $CCR6^{+}$ regulatory T-cells isolated from mice (141) but it was one among several miRNAs and so not focussed on. Besides, Dunay et al have shown that, *in vivo* at least, CCR6 does not mediate any kind of response (positive or negative) to survival after infection with *T. gondii* (142), which makes this a rather undistinguished connection to parasite infection. Another large-scale miRNA study identified mmu-miR-3080-5p as being downregulated during hypoxia in mouse kidneys (143). This may be more relevant to *T. gondi*-host interactions, given that Hypoxia inducible factor 1,

alpha subunit (HIF1A) is known to be modulated following infection (138, 144, 145). In the hypoxia/miRNA experiment, Ho et al found that, strikingly, DICER overexpression in Human Umbilical Vein Endothelial Cells resulted in a reduction in HIF1A following 24h of growth in hypoxic conditions. Similarly, HIF1A-target genes (such as Hexokinase 2) were also downregulated in a Dicer-dependent manner, under hypoxia. However, the situation appears to be more complex: *Dicer* mRNA too was underexpressed under conditions of hypoxia. What this implies for mmu-miR-3080 is extremely unclear, given that it was one of 148 miRNAs to have been downregulated as a consequence: it may simply be a casualty of mass *Dicer* downregulation during a hypoxia-like state induced by parasite-infection, but there is far too little evidence to support this idea at the moment (HIF1A and *T. gondii* infection are discussed in much greater detail in **Chapter 6**).

mmu-miR-200b-3p

The majority of known miR-200b functions are related to the epithelial-mesenchymal transition, a key process for differentiation and metastasis. In the context of cancer, miR-200b has been found to be an antagonist of this transition – this finding extends to several types of cancer and it is considered a tumor suppressor (146). Interestingly, miR-200 family members have also been shown to be transcriptionally upregulated by MYC (147) which is likely to be related to the upregulation of MYC in *T. gondii*-infected cells (21) (this too is discussed in further detail in Chapter 6). Another related function of miR-200b's is its involvement in the cell cycle. Yu et al were exploring monocyte/macrophage differentiation and the downregulation of p38 mitogen-activated protein kinase (MAPK) interacting protein (p38IP) that occurred during it. They searched for potential miRNAs that might target that gene and miR-200b-3p was a candidate. A luciferase assay and target mutagenesis showed that it was a direct target. Crucially, p38IP has been shown to be a

mediator of the G2/M cell cycle transition and downregulation of p38IP blocks this transition. Thus, it could be that an upregulation of mmu-miR-200b works towards blocking the cell cycle at this stage – which is a well-known effect on host cells by *T. gondii* (148, 149).

Despite these two intriguing miRNAs being dysregulated following infection by the Type II strain ME49, the fact that the depth and quality of the sequencing libraries were too low to obtain more information is rather disappointing. That being said, the complexities of NGS analysis are often best done 'in practice' rather than in theory. Subsequent efforts are likely to be of far higher quality – not just because of technological advances in the library preparations and sequencing themselves but also the available algorithms and quite simply my ability to think critically about these different parameters. For that reason, the following chapter demonstrates a more technically-sound analysis of host cell miRNA dysregulation in the face of *T. gondii* infection.

V. A Deeper Examination of microRNAs and *Toxoplasma gondii*

5.1 Introduction

Despite the ultimately unsatisfactory sequencing depth (which thus tainted all further analyses), a few interesting preliminary results did emerge. As a result, I decided that to repeat and extend the experiment to include more samples, using a more robust sequencing platform would be valuable in exploring the extent to which infection by *Toxoplasma gondii* has an effect on host cell microRNA expression.

5.1.1 Normalisation

Given that my data were of much greater depth and quality, I was able to take advantage of better and more robust statistical methods to normalise my data and thus obtain more trustworthy results. While RNASeq analyses are free from the issues of uneven hybridisation and dye-effects that can plague microarrays, the data still require thorough normalisation, to account for variation in library size and composition. Several methods of RNASeq normalisation have been proposed and, in the past few years, compared. The most commonly-used normalisation methods for NGS data are summarised below:

1. Count-per-million: This is perhaps the simplest method of normalisation, where counts are scaled to library size, and library size only.
2. Total Count: The total count method adds a step, by, following the size adjustment, scaling all libraries by a common factor, usually the mean or median library size.

3. Upper-Quartile Scaling: This adjusts for a large number of low- or zero-count tags, by only considering the upper-quartile of non-zero counts, instead of the total count.
4. Median: Rather than the upper-quartile, the median value of tags across samples is used to scale each library's counts
5. DESeq, the normalisation method implemented within the DESeq package (150, 151). First, the data are scaled according to the median of observed:geometric mean ratio for all tags (in all libraries). Following this scaling, the data are modelled according to a negative binomial distribution. This normalisation method relies on the assumption that most tags (genes) will not be differentially-expressed.
6. Trimmed Mean of *M*- values: Like the DESeq method, TMM also assumes that most genes are not differentially-expressed, but, also takes into account the RNA composition of libraries. Robinson et al (152) argue that, while the 'simple' count-based normalisations (1-4 in this list) might be appropriate for scaling between replicates, they are inadequate for looking at biological differences. This is due to what they call "sequencing real estate": the idea that if libraries are sequenced to similar depths, differences in library composition will have a large effect on tag counts. The overcome this, TMM first assigns as "reference samples" those observations whose expression is close to the mean expression of all tags. The remaining samples (the "test" ones) are trimmed of observations with the highest log-fold changes (*M*-value of 30%) and absolute expression (*A*, of 5%) and then scaled. Then, the libraries are scaled based on the weighted mean of test:reference log-ratios. The fact that this method takes into account the fact that some reads will be highly-sequenced (as, for instance let-7 miRNA family members) makes it the most appropriate for miRNA sequencing experiments. TMM is the method implemented in EdgeR (140).

A number of studies (153, 154) have conducted comparative analyses of these (and other, less commonly-used) methods on both real and simulated RNAseq data. While most of the methods, other than Total Count, resulted in somewhat similar stabilisation of library size across samples, DESeq and TMM performed far better in terms of lowering the coefficients of variance in the normalised samples. Importantly, following differential-expression analyses of libraries normliased by these methods, TMM and DESeq came out as clear winners, able to both control for false-discovery but also detect more DE genes.

5.2 Methodology

5.2.1 Experimental Design

I constructed the experimental design so as to correspond to a metabolomic study previously performed in the lab, which compared ^{1}NMR profiles of Uninfected, RH-infected and ME49-infected NIH/3T3 mouse embryonic fibroblasts. These time points and MOIs were chosen in order to, as far as practically possible, keep parasite numbers comparable between strains and time points. The rationale behind this was to avoid cases where observed transcriptional effects might be attributable intracellular parasite numbers rather than to strain difference or the duration of infection. This MOI 'normalization' was confirmed by infecting identical host cell monolayers at an MOI of 1.2 for RH and 3 for ME49. Twenty-four and 43 hours after infection, the monolayers were examined under the microscope to tally the number of infected cells for each condition. [This work was done with Dr Aysha Roohi].

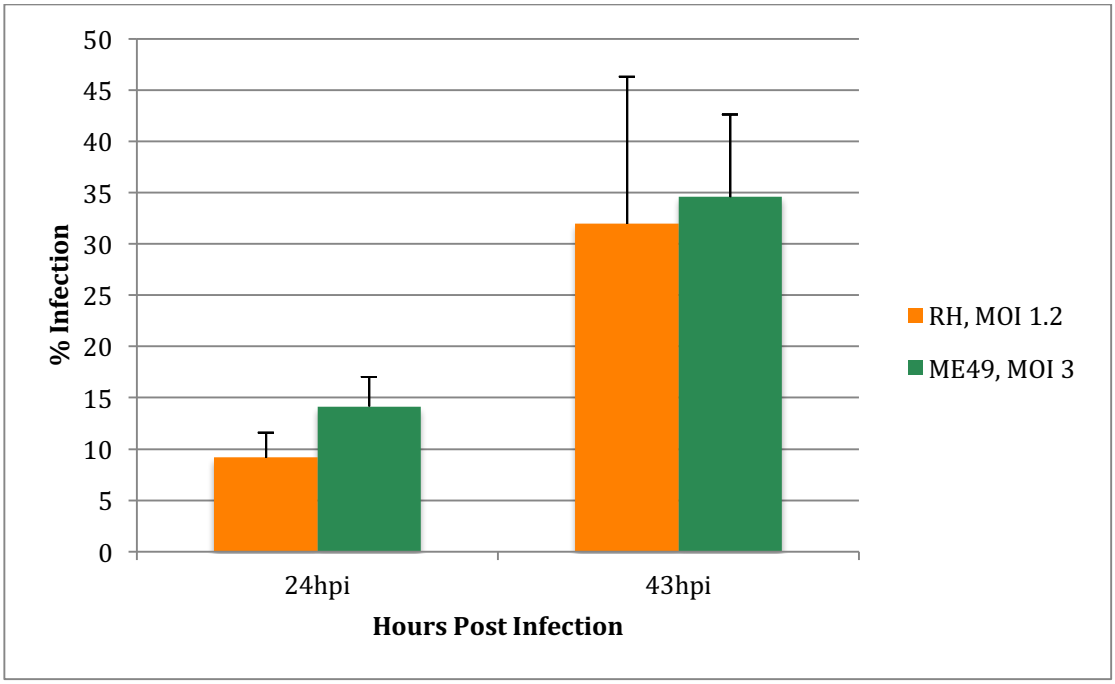


Figure 5.1. Infection Rates with RH and ME49. Cells were infected with either RH or ME49, at an MOI of 1.2 or 3 respectively. After 24 hours, infected cells were counted.

To further guard against transcriptional differences arising from parasite numbers rather than strain-type or infection duration, the MOIs were also applied reciprocally. Ultimately therefore, the time points and MOIs chosen were as in Table 5.1.

Table 5.1. The experimental set-up of infection. Numbers represent replicate wells of host cells either infected or washed with fresh medium at time 0.

Strain	ME49		RH		Uninfected
MOI	1.2	3	1.2	3	NA
0h	0	0	0	0	3
24h	3	3	3	3	3
43h	3	3	3	3	3

NIH/3T3 cells were seeded in 6-well plates, and allowed to adhere and grow for approximately 48 hours (33 sample wells, along with nine wells grown for host cell counting, in order to determine the appropriate number of

parasites to add to achieve the desired MOI). At this point, when host cells were ~70% confluent, purified parasite preparations were made (see **2.3.3**) and the appropriate volume was added, in fresh, supplemented DMEM medium, to each well, so as to achieve the desired MOI. For the uninfected (or, 'mock infected') samples, this meant purely replacing the conditioned medium at the first time point.

5.2.2 Protocol Refinement

At each time point (0hpi, 24hpi and 43hpi), RNA was extracted using the Direct-zol method (described in **2.5**). These extracts were then subjected to microfluidic electrophoresis using a Bioanalyzer 2100. This instrument evaluates the integrity of RNA samples according to four main features, each contributing to that sample's RNA Integrity Number (RIN):

1. 18S/28S:Total RNA Ratio, looking at the ratio of RNA under the 18S and 28S peaks to that under the rest of the plot
2. Height of the 28S peak, whose degradation is the first sign of overall sample degradation
3. 'Fast' area Ratio, looking at the area between the 18S and 5S peaks. Increase in this range indicates early signs of degradation.
4. Large, diffuse peaks at short lengths, indicating severe degradation

A critical feature in any RNA extraction is that of desiccating samples (especially if, as in my case, they are to be shipped abroad for library preparation and sequencing). The two main methods for drying are a) leaving tubes in a hood, or b) using a Speedvac. An initial protocol of leaving extracted RNA in a hood overnight (followed by shipment in RNAStable tubes) showed a great deal of degradation (Figure 5.2).

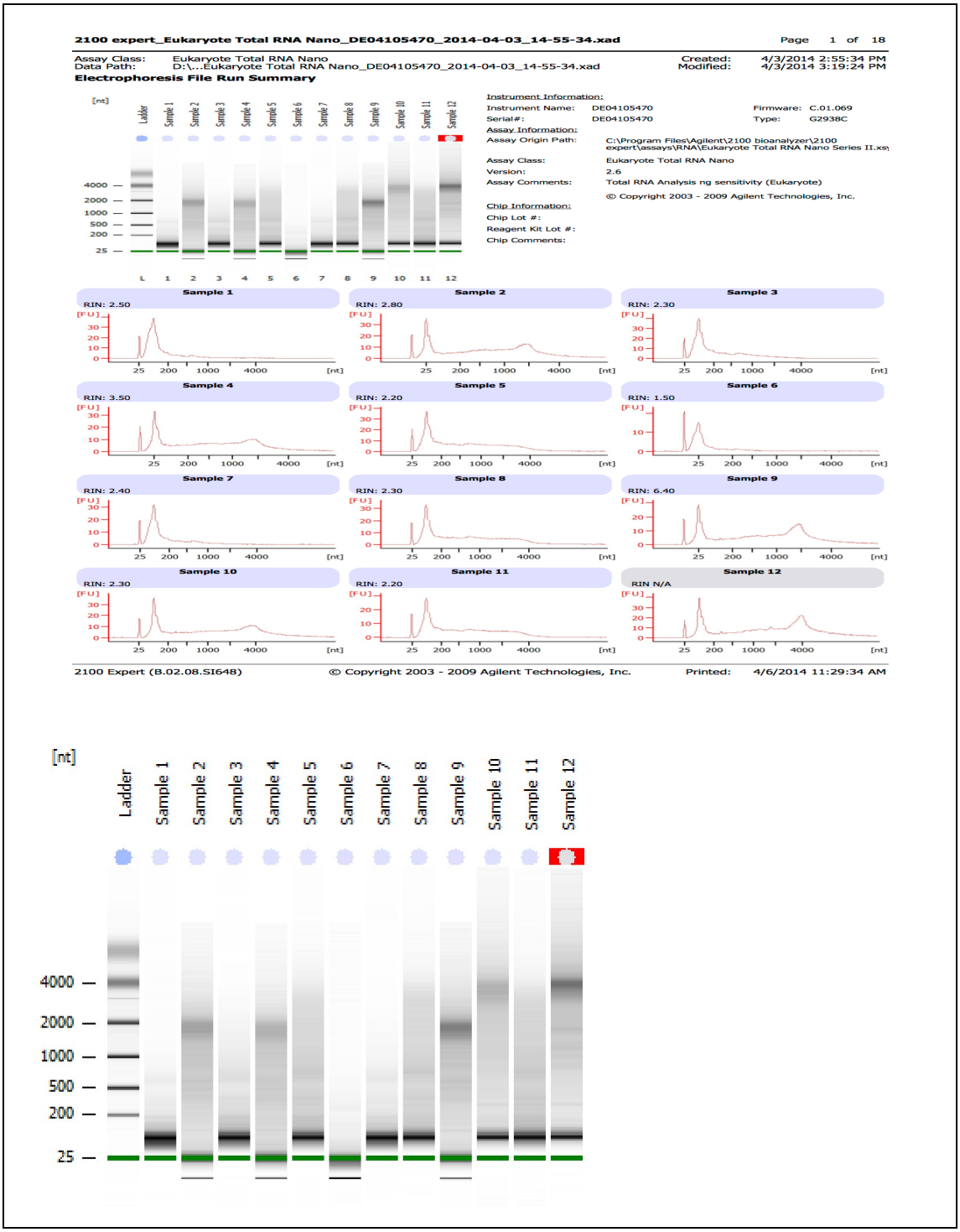


Figure 5.2. Microfluidic Analysis of 12 pilot RNA extracted from uninfected NIH/3T3 cells; Bioanalyzer 2100.

Here, it appeared that, of 12 prepared samples, most had suffered degradation. It was postulated that the samples had not been sufficiently dried or perhaps had RNAses introduced during the drying period. As such, a

pilot experiment was designed to explore the impact of different drying methods on extracted RNA.

Trizol-stored lysates from two biological replicates of NIH/3T3 cells grown to ~70% confluence in 6-well plates were used for this pilot analysis. RNA from these samples was extracte using the Directo-zol protocol, until the elution step (in 60uL RNAse-free water) each. At this point, the samples were split into 3 and treated as shown in Figure 5.3.

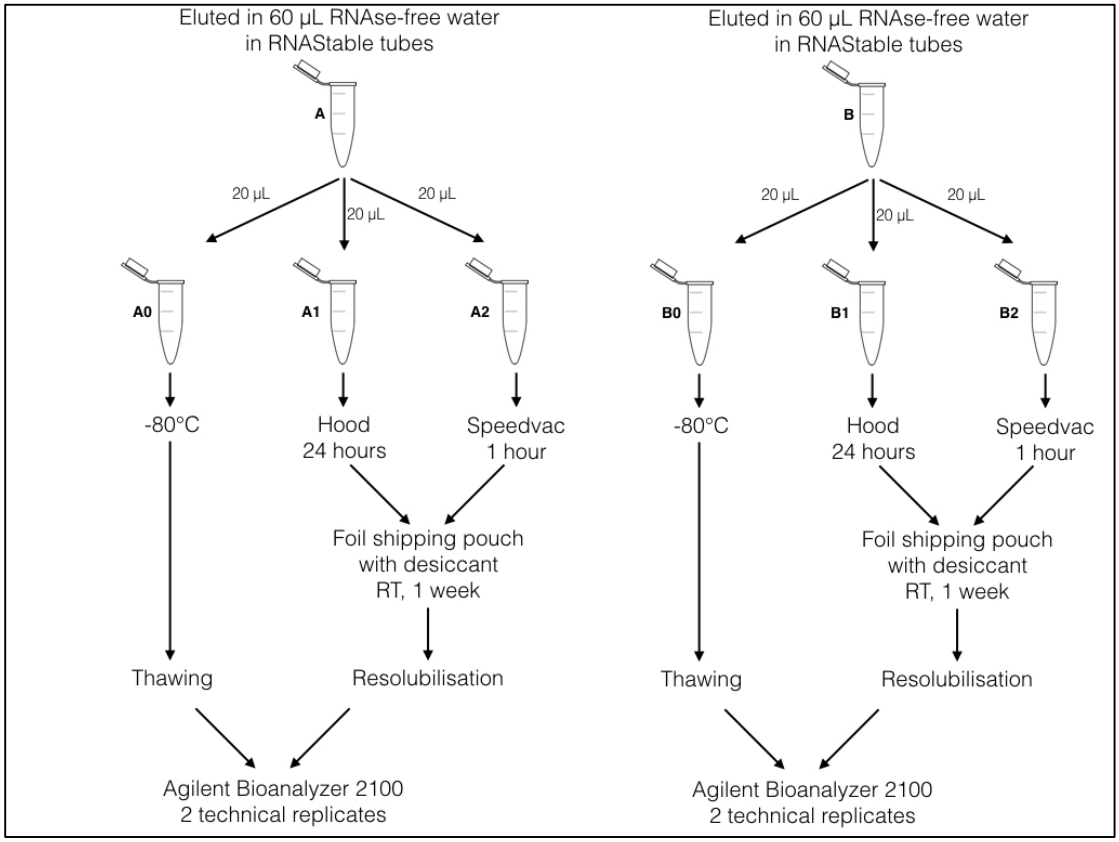


Figure 5.3. Experimental Scheme to examine the effect of drying on RNA quality

The RNA samples were treated according to different recommended drying protocols (as recommended by the manufacturers of the RNAStable storage and shipping tubes). The treated samples were analysed on an Agilent Bioanalyzer 2100 for RNA integrity.

Two RNA Integrity analyses were performed per biological replicate, under the three storage/dessication regimes, yielding 12 analyses in total, as shown in Figure 5.4.

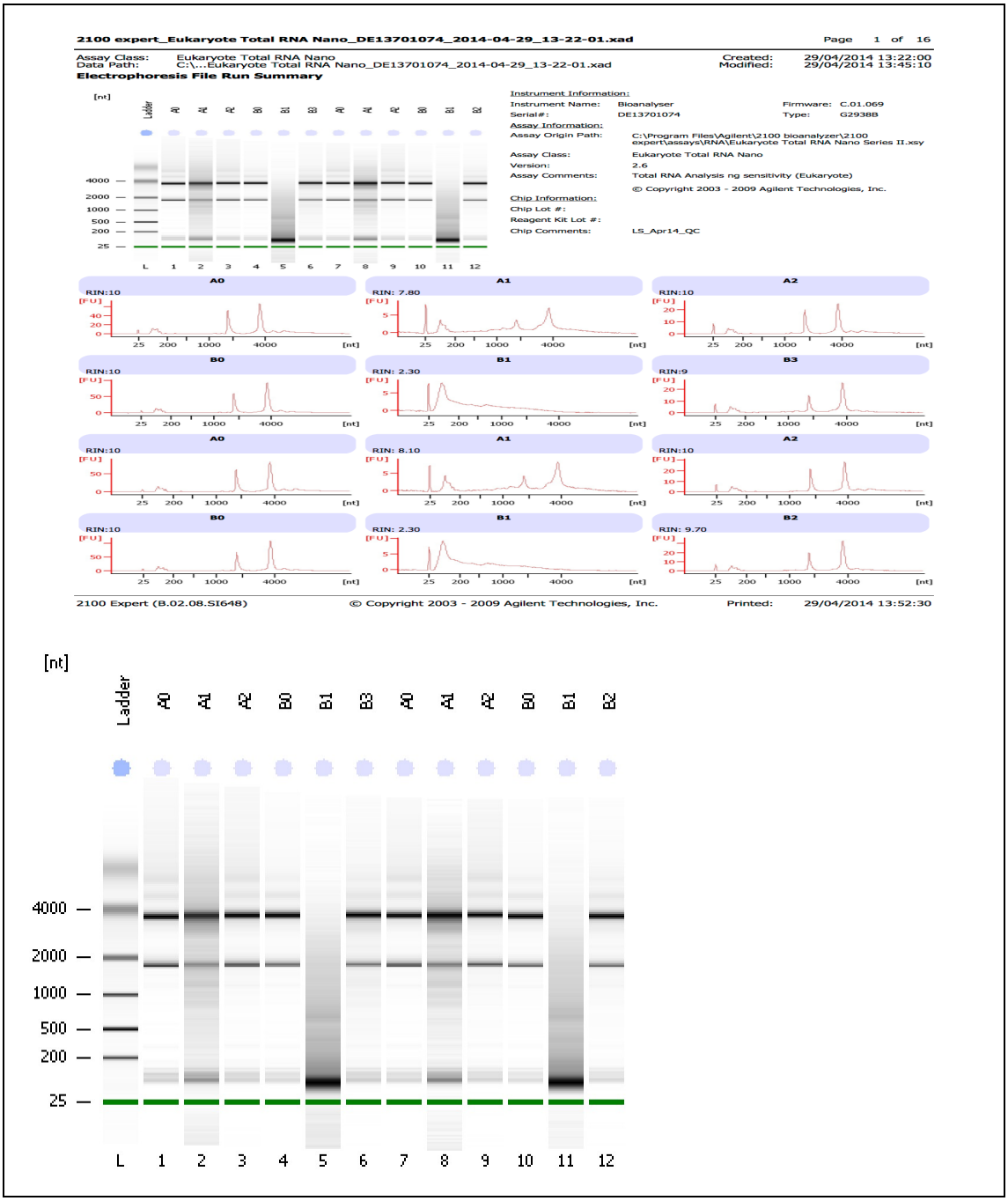


Figure 5.4. Microfluidic Analysis of the 'drying pilot' RNA samples as described in Figure 5.3.

A drop in RNA Integrity (as measured by the RIN value) is apparent in all cases, as compared to the immediately-frozen (never dessicated) sample. In the first biological sample (A), dessication using the Speedvac resulted in no reduction in RIN value, and only a very modest reduction in RNA concentration (mean reduction by 6%), while overnight drying in a hood performed worse, resulting in reductions in both the RIN value (by 2.05

points) and RNA concentration (mean reduction by 7.1%). In Biological Sample B, the effect of the Speedvac is similar, resulting only in modest RIN value reduction (by only 0.65 points) and loss of RNA (9.7%). The effect of hood-drying on Sample B's RNA concentration is an undramatic 13.7% but the drop in RNA quality is precipitous: 7.7 RIN points. This is supported when looking at the electropherograms, where after hood drying, the clear 18S and 28S rRNA peaks disappear completely, revealing instead an overabundance of degraded RNA (large low-MW smear),

5.2.3 RNA Extractions and Quality

As a result of the potential variability in using hood-drying as a means of RNA dessication, the Speedvac protocol was used for all subsequent RNA extractions (that is, the 33 samples as described in Table 5.1). A summary of the integrity data (after drying and shipping) of the 33 samples is presented in Figure 5.5.

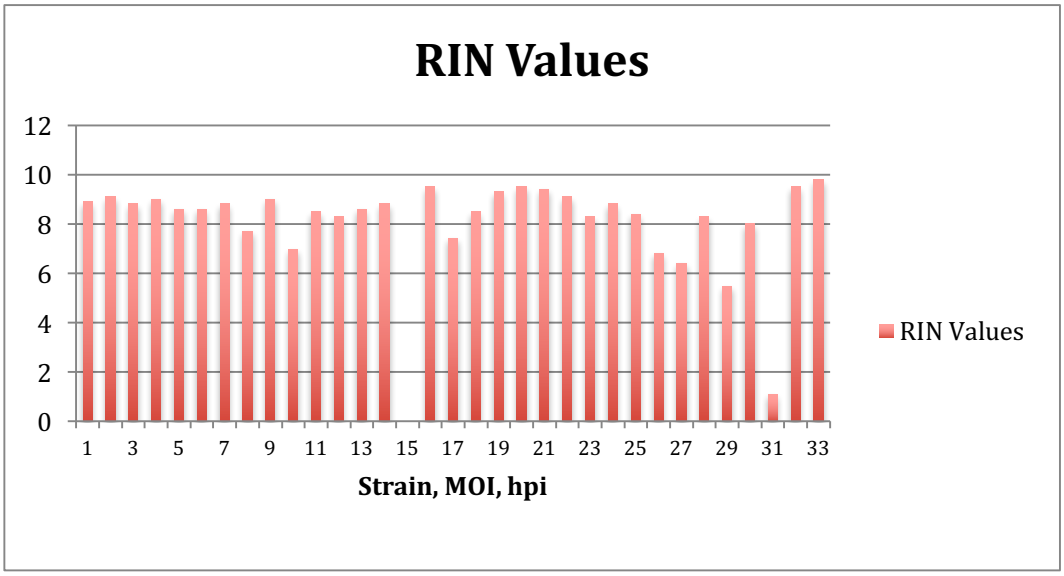


Figure 5.5. RNA Integrity (RIN) Values of the 33 RNA samples.

1-3: ME49 MOI 3 24hpi; 4-6: ME49 MOI 1.2 24hpi; 7-9: RH MOI 3 24hpi;
10-12: RH MOI 1.2 24hpi; 13-15: Uninfected 24h; 16-18: ME49 MOI 3 43hpi;
19-21: ME49 MOI 1.2 43hpi; 22-24: RH MOI 3 43hpi; 25-27: RH MOI 1.2 43hpi;
28-30: Uninfected 43h; 31-33: Uninfected 0h

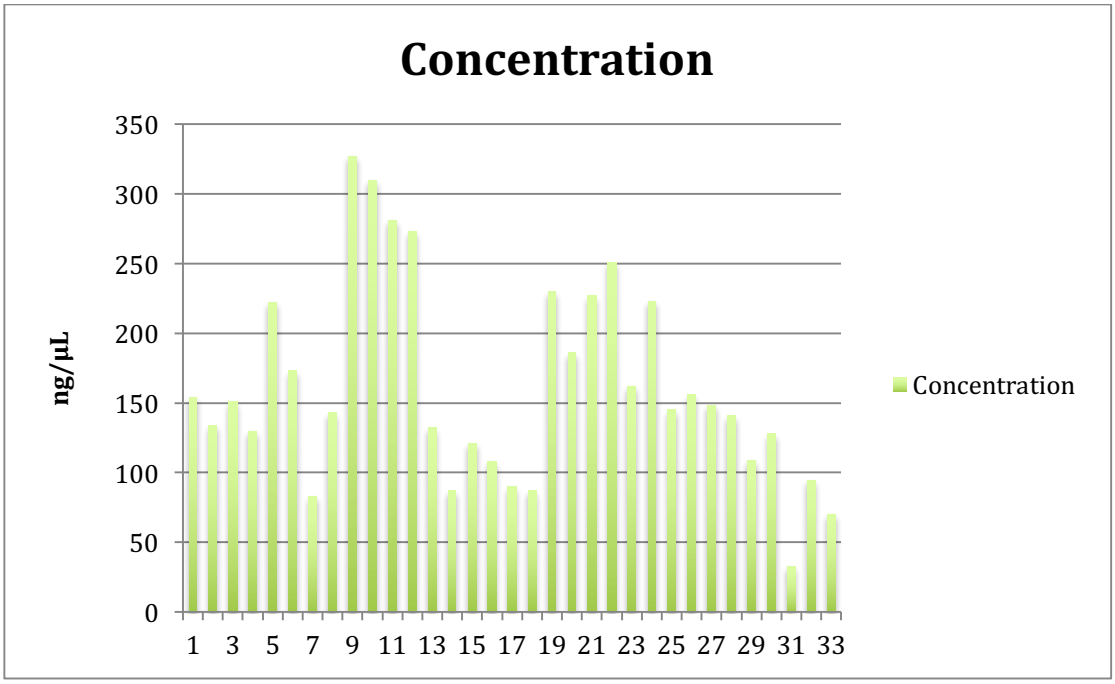


Figure 5.6. Concentration of the 33 RNA samples

1-3: ME49 MOI 3 24hpi; 4-6: ME49 MOI 1.2 24hpi; 7-9: RH MOI 3 24hpi;
10-12: RH MOI 1.2 24hpi; 13-15: Uninfected 24h; 16-18: ME49 MOI 3 43hpi;
19-21: ME49 MOI 1.2 43hpi; 22-24: RH MOI 3 43hpi; 25-27: RH MOI 1.2 43hpi;
28-30: Uninfected 43h; 31-33: Uninfected 0h

While no standard cut-off exists for using RINs to decide whether to exclude or re-extract RNA samples, the higher the score, the more intact the sample. But, the RIN must be taken together with the electropherogram traces. Three samples stand out in these analyses: Samples 15, 29 and 31. In Sample 15, while distinct peaks can be seem at the 18S and 28S rRNA regions, an unexpected signal also occurs at the 5S region (~120nt). The presence of this peak triggers Agilent's critical anomaly feature, which prevents the RIN from being calculated. However, the clear presence of the 28S and 18S peaks indicates that this sample should still be usable. On the other hand, it is less clear whether Samples 29 and 31 are still intact enough for downstream processing. While sample 29 has a visible peak at, 18S, no distinct peak can be seen at 28S (hence the rRNA ratio of 0). For sample 31, the situation is worse still, with no distinct peaks, bar the marker, visible at

all. That being said, these samples were still sequenced along with the others, pending exclusion from downstream processing, depending on what the sequencing results yielded. Luckily, these samples were also part of distinct replicate groups (Sample 29 being part one of the three 'uninfected, 43hpi' samples and Sample 31 part of the 'uninfected, 0hpi' group) and their eventual possible exclusion should still make differential expression analysis possible.

Libraries were prepared from the 33 RNA samples, using Illumina's TruSeq Small RNA Kit. [Library preparations and sequencing were performed by Mr Abhinay Ramaprasad, at the Pathogen Genomics Laboratory, KAUST].

5.2.4 Functional Analysis

Given that many of the current enrichment analyses tools do not include miRNAs at all, or include them in a very limited capacity, the usual method that is followed is to generate lists of putative miRNA targets for each dysregulated miRNA and then perform enrichment analyses on these (KEGG or GO, for instance). I opted to use targets from the miRNet database (155), which aggregates experimentally-validated miRNA targets from a variety of large-scale target-finding experiments (such as HIgh-Throughput Sequencing of RNAs isolated by CrossLinking ImmunoPrecipitation, HITS-CLIP experiments) for my lists of dysregulated miRNAs. These experimentally-validated target genes were then the input for use on the online tool WebGestalt (WEB-based GEne SeT Analysis Toolkit) (156, 157). I used as my background the list of experimentally-validated targets from all the miRNAs currently in miRBase version 21 (80).

5.3 Results

5.3.1 Quality of the Sequenced Libraries

Characteristics of the sequenced libraries are presented in the tables below. Read counts are in Figure 5.7 and the collapsed read counts are in Figure 5.8. It is not an unusual feature of RNASeq for libraries to be of quite different sizes. As expected, the profile of read counts when collapsed is similar to that of the total reads and this can be seen in what I term the 'diversity factor' in Figure 5.9.

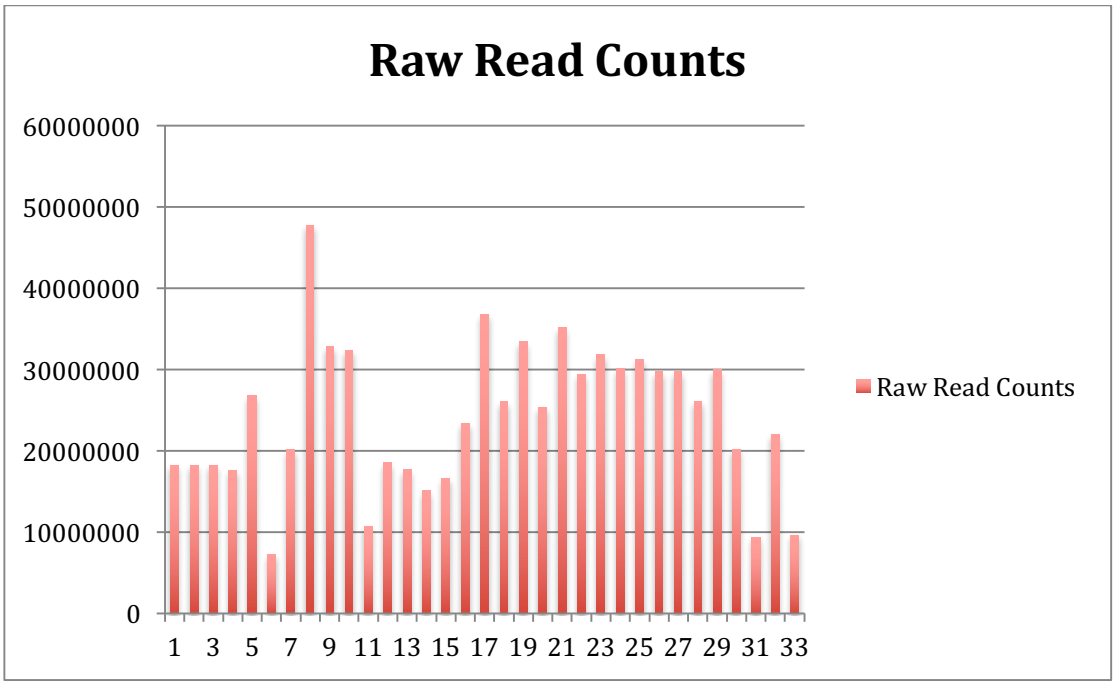


Figure 5.7. Number of reads generated by the sequencer, per sample

1-3: ME49 MOI 3 24hpi; 4-6: ME49 MOI 1.2 24hpi; 7-9: RH MOI 3 24hpi;
10-12: RH MOI 1.2 24hpi; 13-15: Uninfected 24h; 16-18: ME49 MOI 3 43hpi;
19-21: ME49 MOI 1.2 43hpi; 22-24: RH MOI 3 43hpi; 25-27: RH MOI 1.2 43hpi;
28-30: Uninfected 43h; 31-33: Uninfected 0h

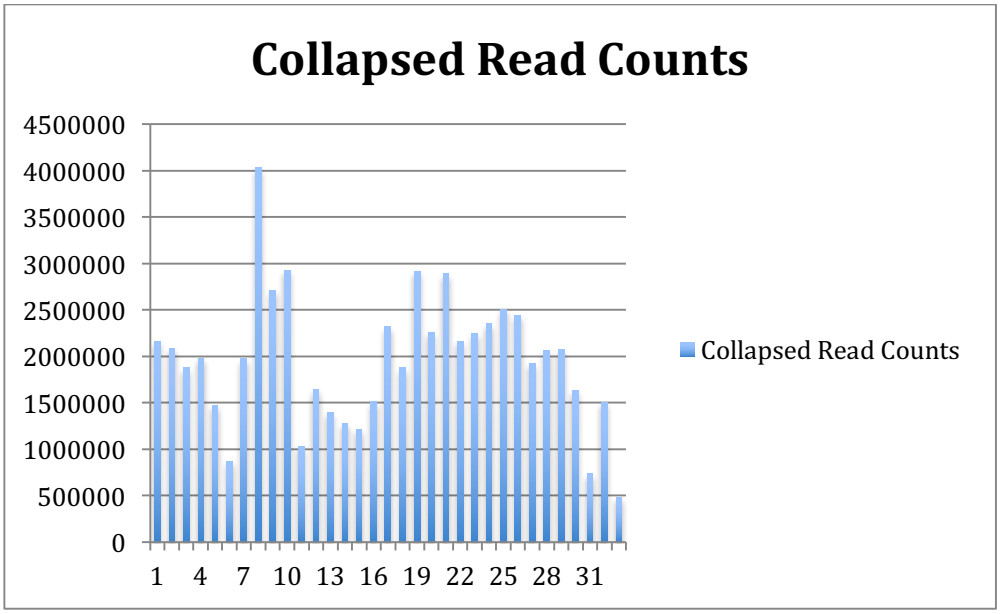


Figure 5.8. Number of 'unique' reads per sample

1-3: ME49 MOI 3 24hpi; 4-6: ME49 MOI 1.2 24hpi; 7-9: RH MOI 3 24hpi;

10-12: RH MOI 1.2 24hpi; 13-15: Uninfected 24h; 16-18: ME49 MOI 3 43hpi;

19-21: ME49 MOI 1.2 43hpi; 22-24: RH MOI 3 43hpi; 25-27: RH MOI 1.2 43hpi;

28-30: Uninfected 43h; 31-33: Uninfected 0h

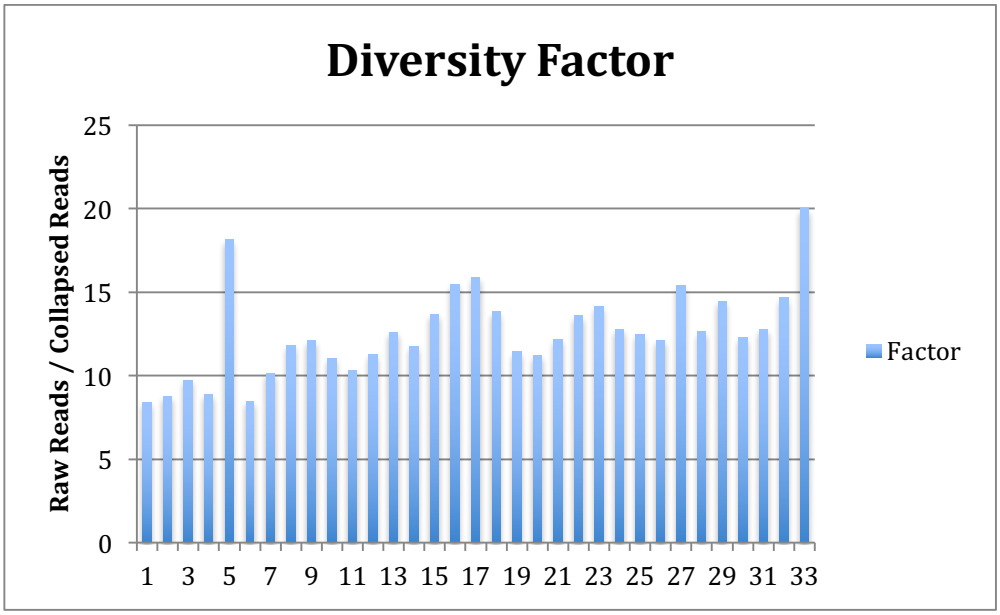


Figure 5.9. The ratio of collapsed to raw reads, per sample

1-3: ME49 MOI 3 24hpi; 4-6: ME49 MOI 1.2 24hpi; 7-9: RH MOI 3 24hpi;

10-12: RH MOI 1.2 24hpi; 13-15: Uninfected 24h; 16-18: ME49 MOI 3 43hpi;

19-21: ME49 MOI 1.2 43hpi; 22-24: RH MOI 3 43hpi; 25-27: RH MOI 1.2 43hpi;

28-30: Uninfected 43h; 31-33: Uninfected 0h

I then aggregated the 33 libraries of raw sequencing reads and subjected them to quality control using FASTQC (158). This tool is aimed at identifying any potential issues with the libraries' qualities, whether arising from problems with the library preparation or from the sequencing itself.

Per base sequence quality

The threshold for sufficient quality is depicted in Figure 5.10 as the green box. Ordinarily, this quality check shows a box-and-whiskers plot of each base position along the read. Our sequences are of extremely high quality – all well within the 'very good' range of Phred scores above 28. The overall mean value is of 36.2. As indicated in Fig, the first five bases are of slightly lower quality (with only the first five bases having a mean quality of 32.8 as compared to a mean of 36.6 if these five are excluded). This value still places them well-above the threshold for acceptable quality, and is likely due to the presence of barcodes. As has been noted, a lower start-of-read quality does not impinge on the usability of the reads and is likely due to a historical artefact of Illumina's own quality control procedures (159).

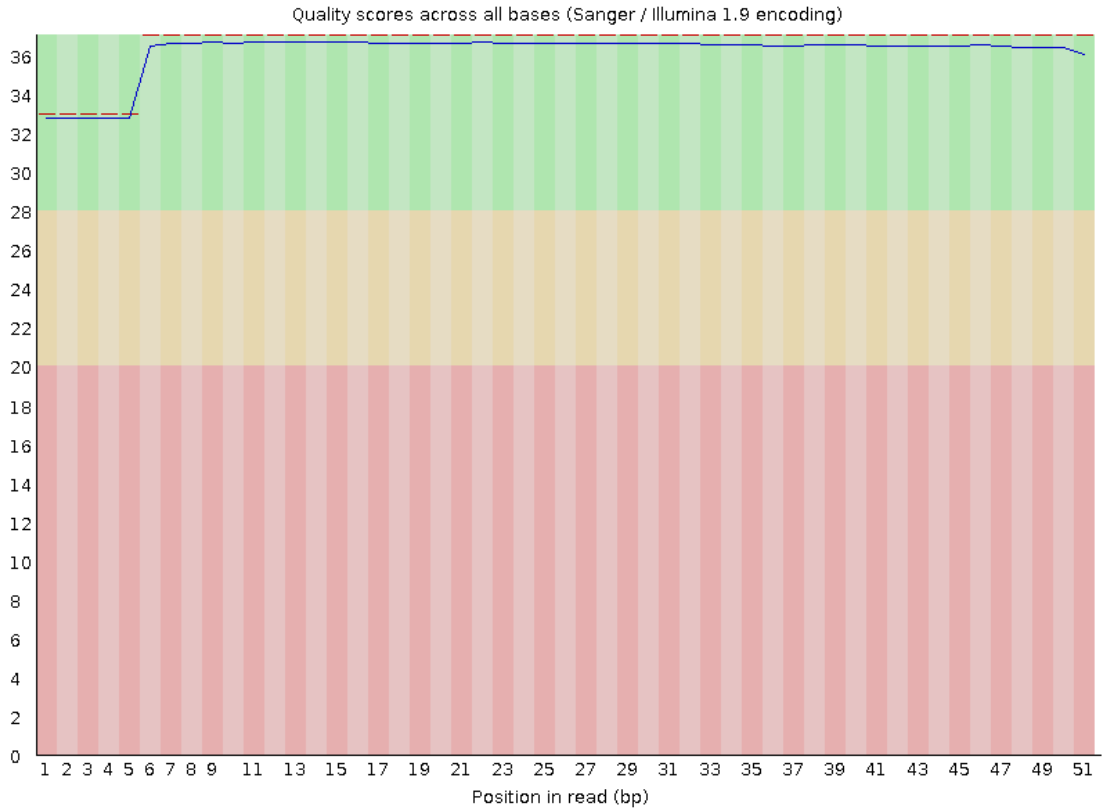


Figure 5.10. Phred quality of all libraries taken as a batch, across read length

On an individual level, most of the libraries exhibited this pattern, with mean and median qualities as well as lower whiskers still being well-within the highest (green) range. Two exceptions were 8913 and 8918, where lower-whiskers dipped below the Phred high-quality threshold of 28. However, even for these libraries, the mean, median and inter-quartile quality range were all above this level, with only the lower-whisker reaching 27.

Per sequence quality scores

This analysis module computes the average quality per read, to indicate cases where subsets of reads have lower than average quality. In our case, the quality score distribution across the read population is excellent, with most reads having a high average quality, Figure 5.11.

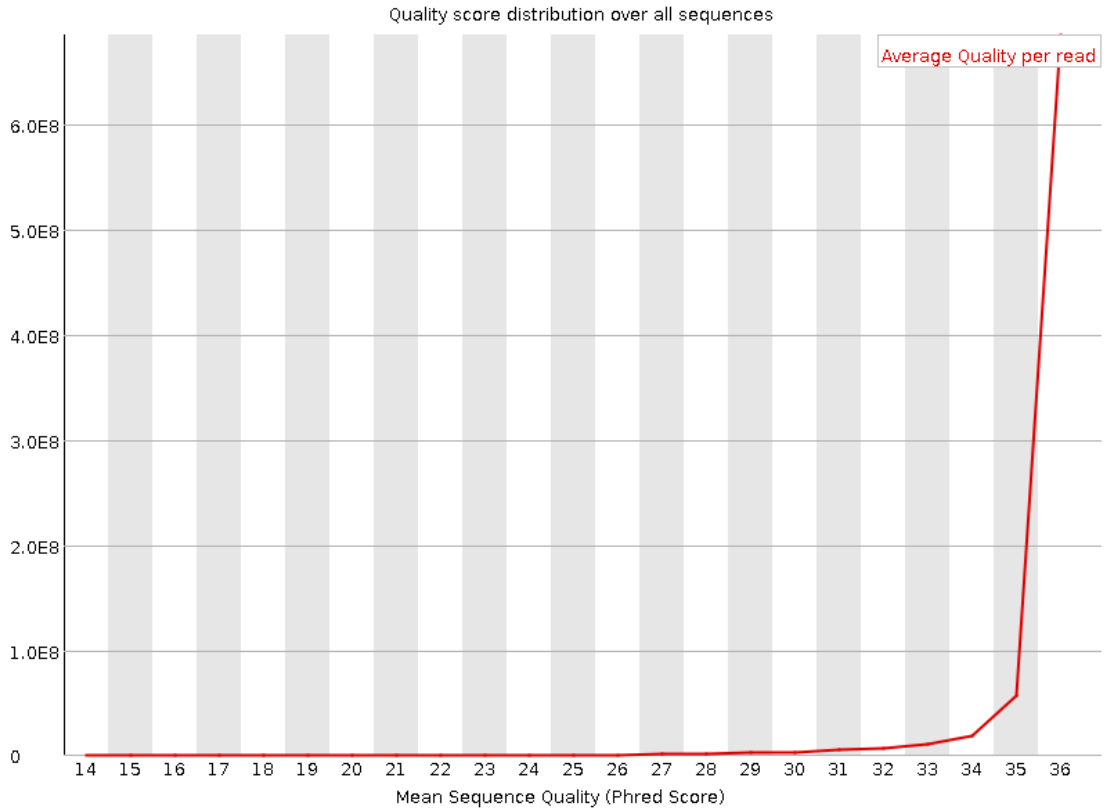


Figure 5.11 – Per-Read quality, across all libraries

This was also the case when looking at the libraries on an individual basis, with no exceptions.

Per base N content

This parameter flags instances where the the sequencer was unable to call a base, thus substituting "N" for the call. Indicative of successful base calls across the entirety of the reads is our extremely low levels of ambiguously-called bases, <= 5% at any position (Figure 5.12). Similarly, the individual libraries also had extremely low levels of ambiguously-called bases.

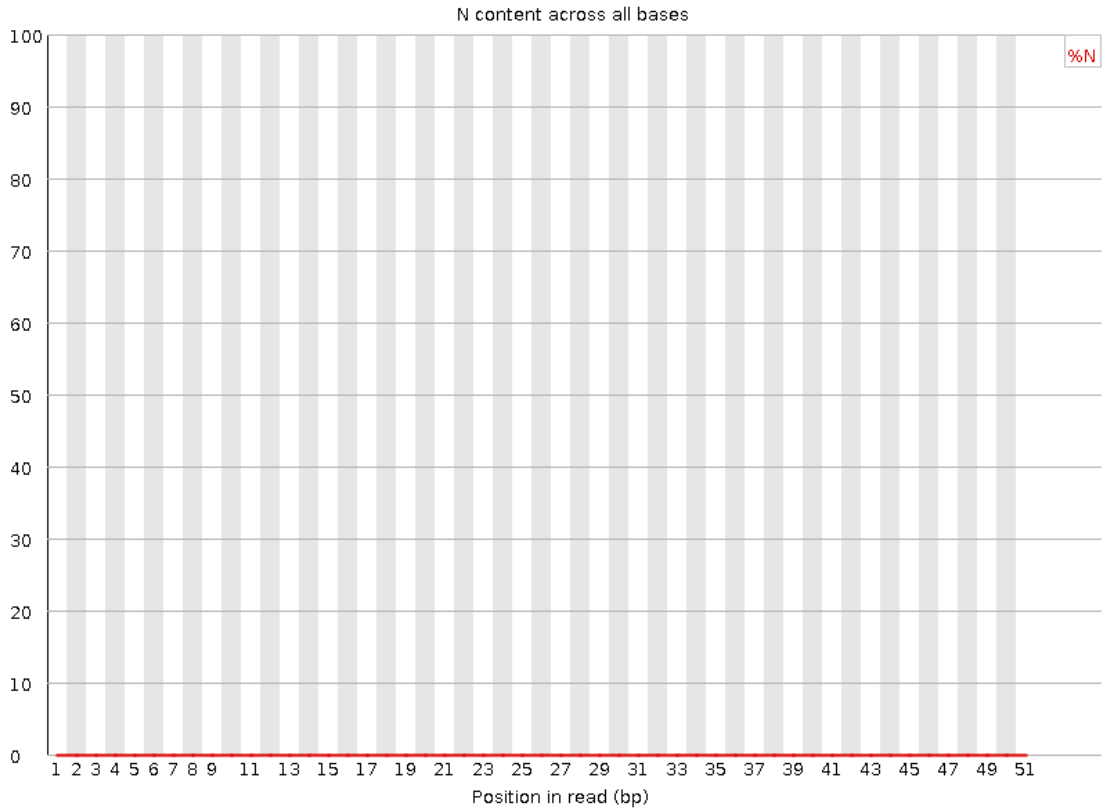


Figure 5.12. Number of ambiguously-assigned bases, over all libraries

Sequence Length Distribution

This graph, with its single peak, indicates that all reads were of the expected size. This is obviously the expected result for libraries where the adaptors have not yet been trimmed, Figure 5.13.

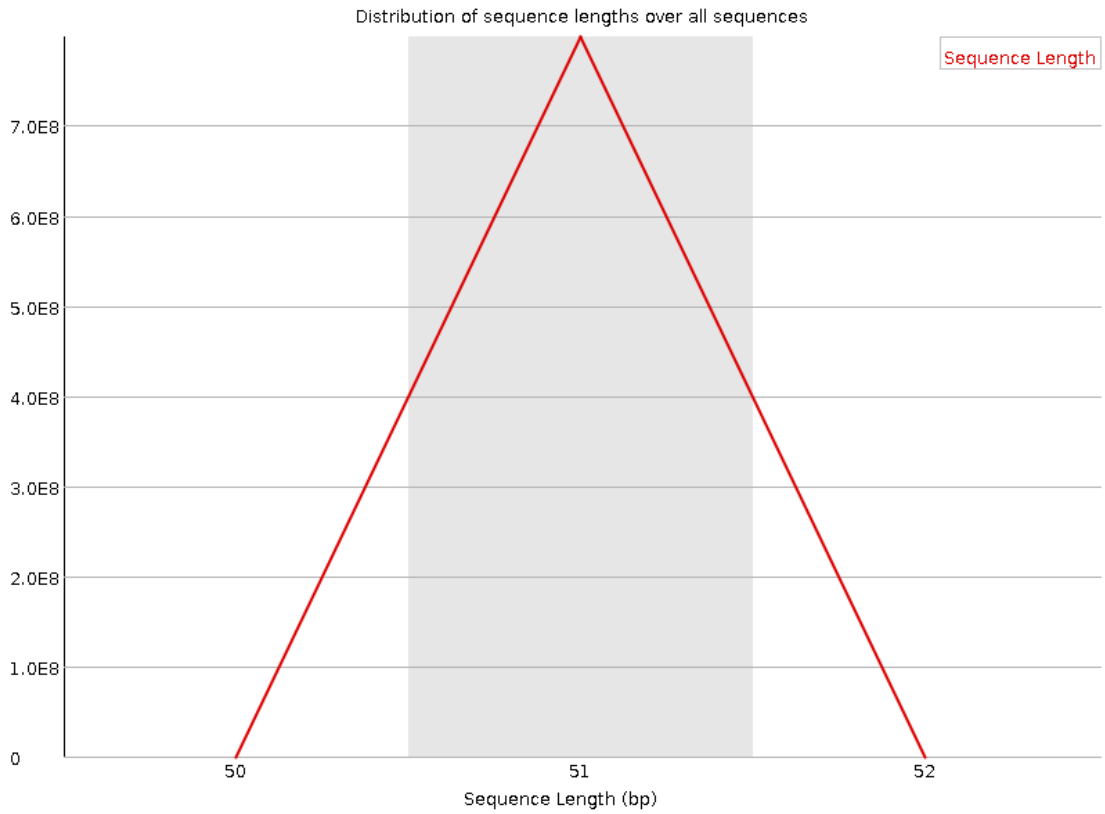


Figure 5.13. Distribution of read lengths, for all libraries

Per base sequence content

In the ideal scenario, the base content for each of the 4 bases would be constant across the reads, according to the genomic distribution. This quality control parameter can be disregarded for small RNA sequencing, as it was designed with whole-genome libraries in mind. A library selected specifically for small RNAs will inherently have biased, non-uniform distributions of nucleotides across the read length. Moreover, the presence of barcodes and adaptors further confuses the issue, contributing to inconsistent traces, Figure 5.14.

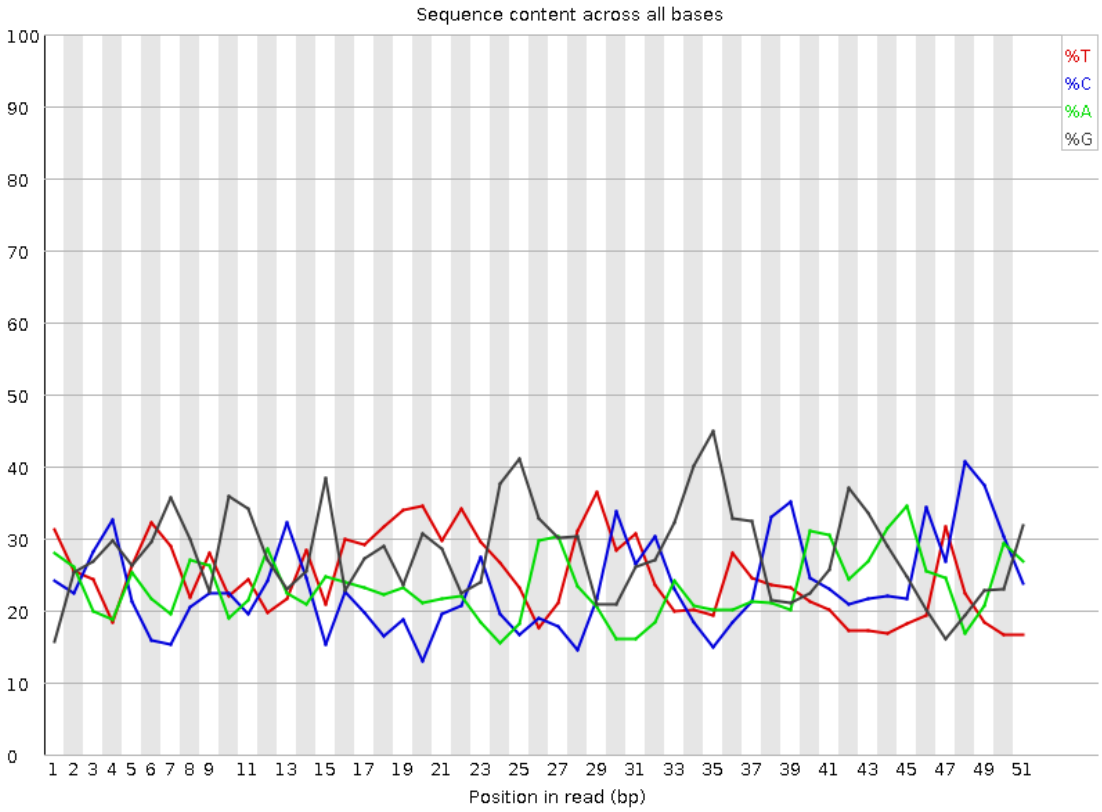


Figure 5.14. A/T/C/G content, per read position, for all libraries

Per sequence GC content

As before, deviation of the GC content from the expected normal distribution is a result of the inherent bias in small RNA library preparation as well as the inclusion of adaptor sequences.

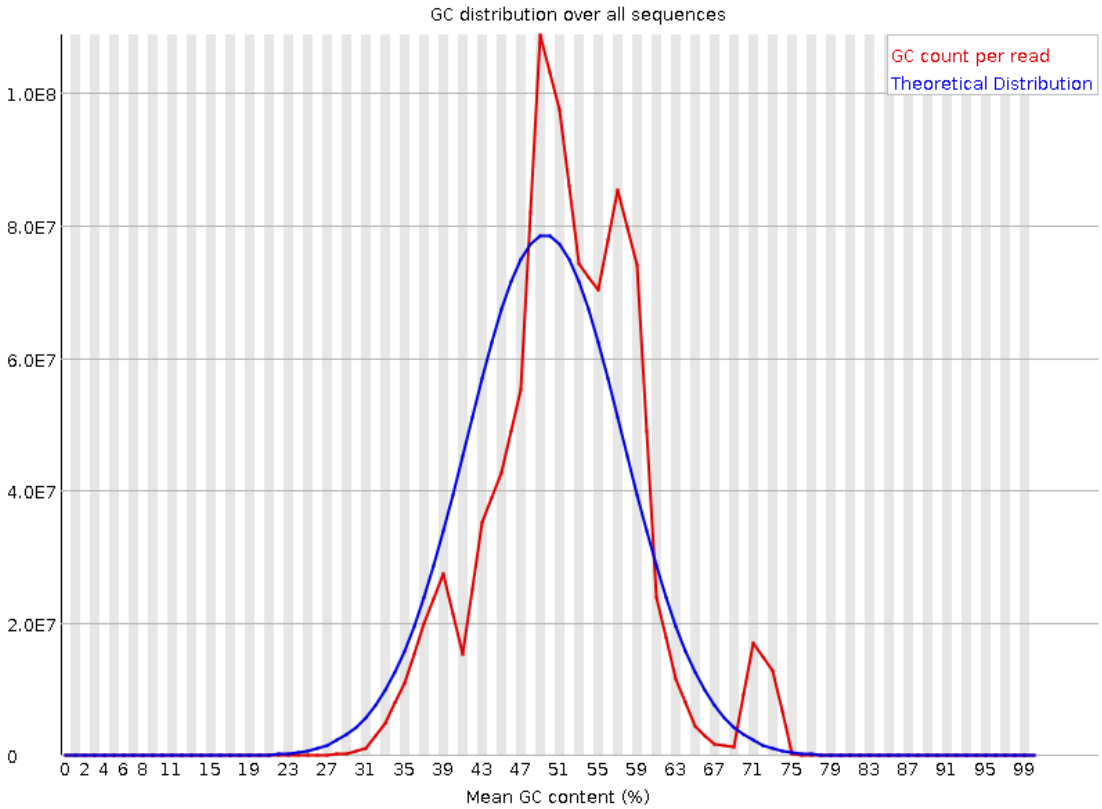


Figure 5.15 – GC-content for all libraries

Sequence Duplication Levels; Overrepresented Sequences; Kmer Content

These parameters are all geared towards genomic libraries, which are not expected to either contain highly enriched sequences (be these full sequences or shorter Kmer fragments) or adaptors. A small RNA library is expected to have skewed distributions of both of these, and the distributions can be accounted for when looking at the FASTQC parameter that now follows.

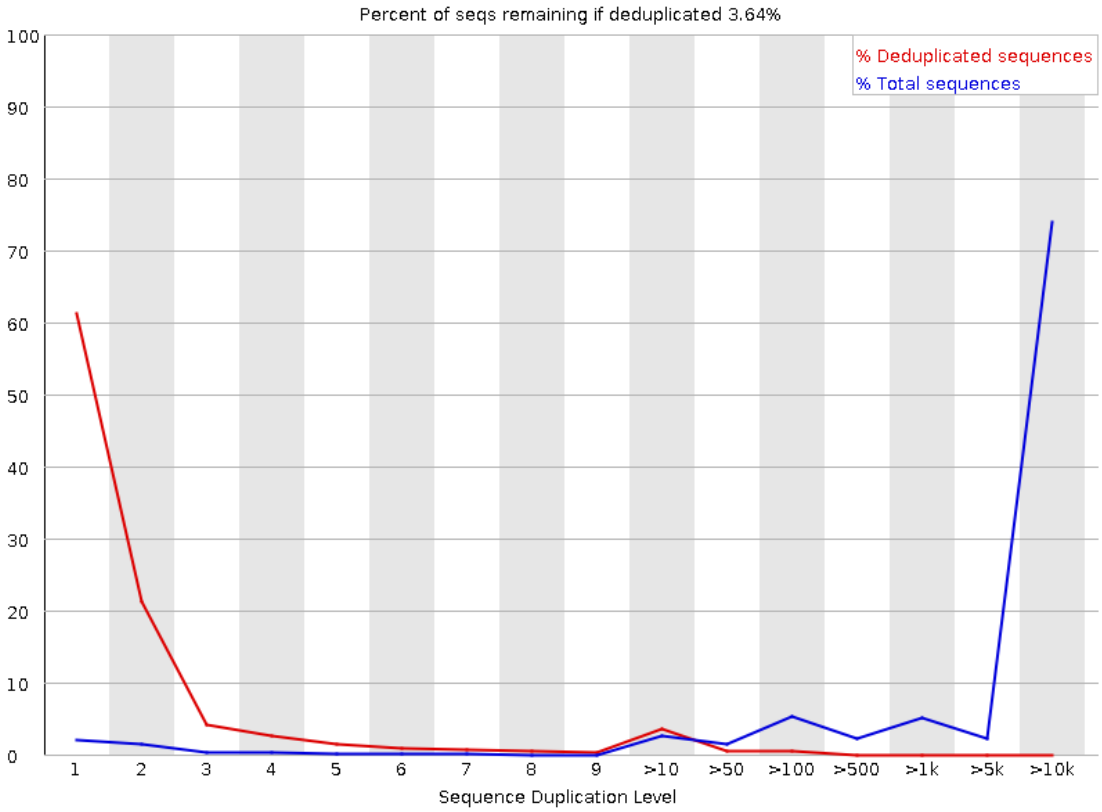


Figure 5.16 – Sequence duplication, for all libraries

Overrepresented Sequences.

As expected with adaptor-containing, size-selected libraries (such as a small RNA library), a number of overrepresented sequences appear in this analysis. In fact, 112 such sequences were highlighted, with 59 of them accounted for as being Illumina primers and adaptors. Of the remaining, many of these fall within rRNA regions, again, as is expected from small RNA libraries. There were also a number of sequences (the most highly-overrepresented one, for example) mapping to *T. gondii*.

5.3.2 Preprocessing and Alignment

5.3.2.1 Preprocessing

The primary tool that I used for pre-processing and alignment of the libraries to the mouse genome was miRDeep2 (113), a tool that returns both quantified known miRNAs (as per the latest release of miRBase, v. 21) and detects novel ones. The miRDeep2 package is a multi-step tool, the usual workflow of which is as follows:

1. Trimming of 3' adaptors
2. Collapsing of identical (trimmed) reads
3. Alignment of the (collapsed, trimmed) to a supplied genome.

While convenient as a 'one stop shop' for miRNA analysis, some of the (immutable) parameters applied by miRDeep2 in two of these steps are suboptimal, namely, adaptor trimming and alignment.

Adaptor Trimming

Within miRDeep2, only the 3' adaptor is searched for and clipped. As seen both in the FASTQ results (where several other Illumina library preparation and sequencing adaptors appear among overrepresented sequences) and in the pilot study in **Chapter 3**, adaptor-contamination, including primer-dimers, can occur at other positions and with not just the 3' adaptor. Moreover, within miRDeep2's trimming algorithm, reads that do not contain any adaptor sequence at their distal end are "retained but not clipped" (129). Given 50nt read lengths, as with my 33 libraries, it is reasonable to assume that if a miRNA is contained within a particular read-insert, it will be contained in its entirety. As a result, the absence of 3' adaptor following the miRNA is likely to represent not a true miRNA but rather a product of mRNA degradation or contamination with rRNA. To prevent the inclusion of such species within my sample, I therefore used a stand-alone adaptor trimming programme, cutadapt

(version 1.9.dev6) (125). This tool offers a number of options for read trimming, including 5' and 3' adaptors, as well as dimers thereof, at any location within a read. Moreover, a possible included option is to require the presence of a 3' adaptor (following the miRNA) in order for a read to be retained. The results of the trimming process are shown in Figure 5.17.

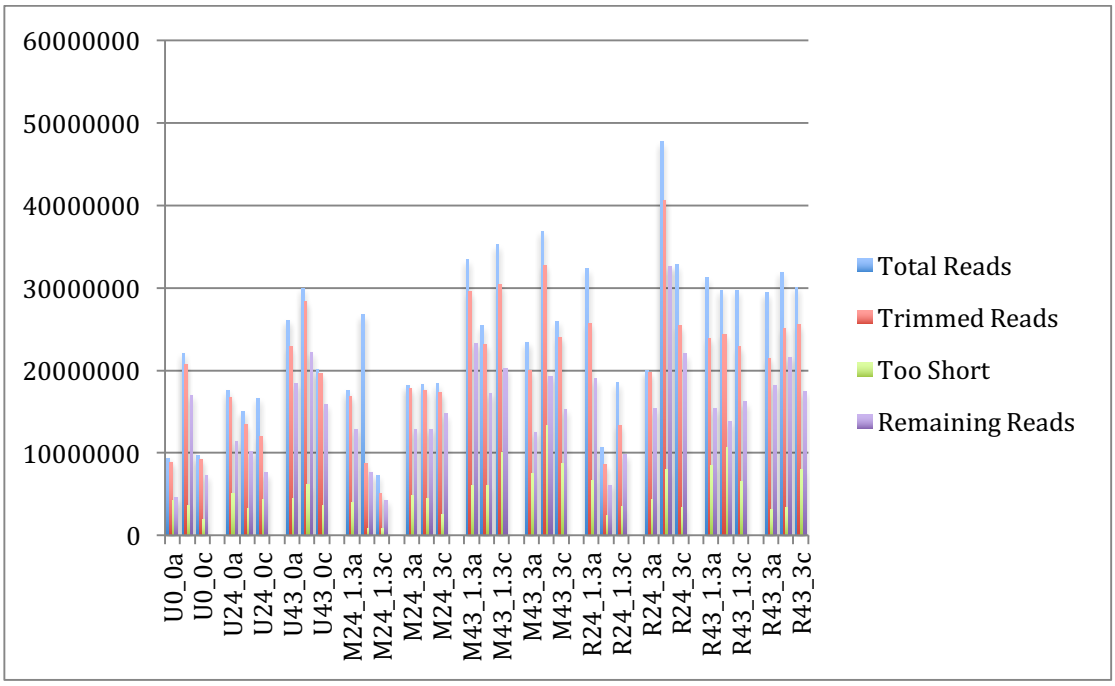


Figure 5.17. Uncollapsed ('raw') reads and the effect of trimming.

Total Reads: The number of raw reads

Trimmed Reads: The number of raw reads that contained the 3' adaptor

Too Short: Of the Trimmed Reads, the number of reads that were smaller than 18nt

Remaining Reads: Reads used in downstream analysis

After trimming, I checked the aggregate quality of the trimmed reads once again, using FASTQC. The overall per base and per base sequence quality remained high. Indeed, the only significantly changed FASTQC reports were, as expected, the sequence length distribution which now had a peak at 22nt (Figure 17), and the most over-represented sequences, which now no longer contained recognisable Illumina primer or adaptor sequences.

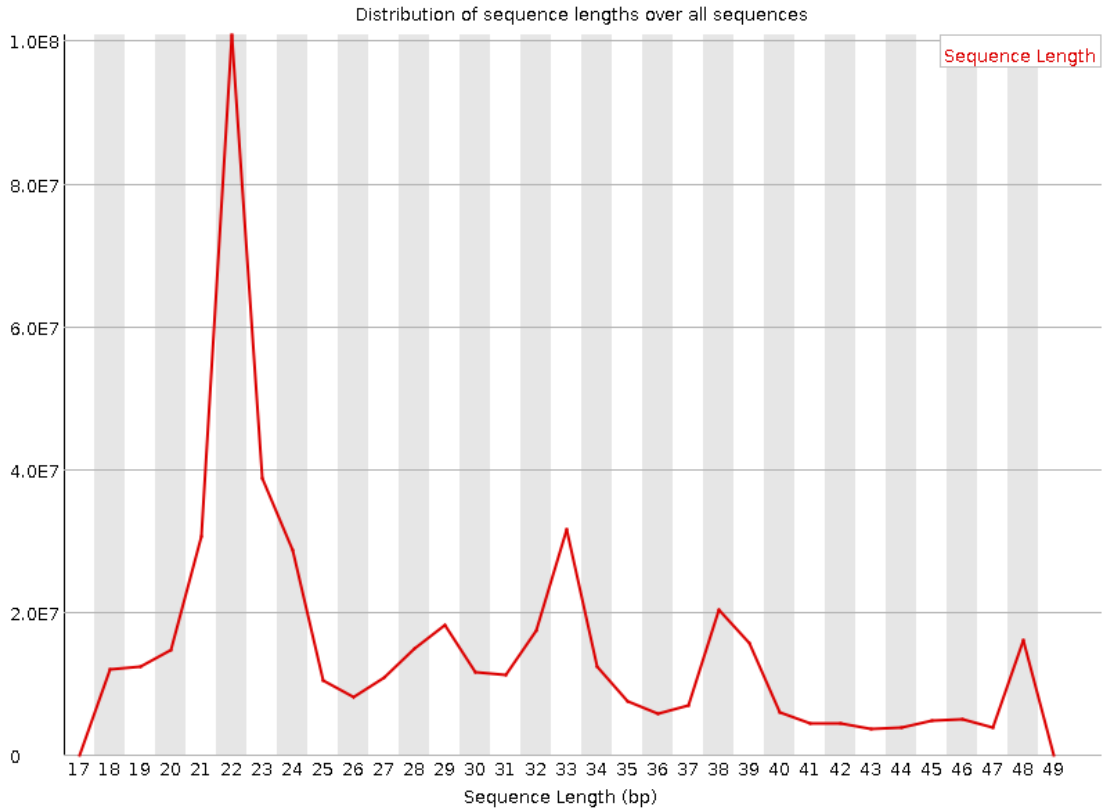


Figure 5.18. Distribution of read lengths after adaptor trimming, for all libraries

Alignment

miRDeep2 uses Bowtie to map trimmed and processed to the supplied genome. Of the numerous variable parameters permitted by Bowtie, only a few are modifiable within the miRDeep2 instance: whether or not to allow mismatches in the seed region and the length of the seed region. Most of the defaults set by miRDeep2 are as required for miRNA alignment, but one key restriction is of concern. As discussed in Chapter 3, in theory, there should be no limit to the number of times a read is 'allowed' to map to genomic locations, and miRNA-seq projects that limit this parameter often do so with little-to-no justification of the cut-off. That being said, the desirable alternative – reporting *all* legal alignments – is not possible within the instantiation of Bowtie implemented by miRDeep2, which requires instead that the user apply a discrete cut-off (i.e. if a read maps more than n times, it is to be discarded).

In order to examine empirically this parameter, I thus first performed an alignment of a single library, using the standalone version of Bowtie (118), allowing for all alignments to be reported.

The largest number of mappings for a single read was 1,678,560 which, as expected, corresponded to a very low-complexity read (GAGAGAGAGAGAGAGAGAGA). Similar, dinucleotide runs emerged from several other extremely highly multi-mapping reads. While such reads would almost certainly be excluded from downstream analyses, their inclusion vastly increases the computing power necessary for these analyses and so, filtering them out at the outset is desirable. That being said, such extreme cases are rare: only 23 reads (of 7,144,900 trimmed reads that aligned anywhere) mapped to over 100,000 locations, with 10,532 (0.15% of aligned reads) mapping to more than 100 distinct locations. Using this cut-off (100 locations) did not noticeably increase the time taken for alignment but halved the rate of suppression due to multi-mapping. As a result, for the miRDeep2-wrapped mapping step, I applied a cutoff of 100 possible mapping locations for each read.

With the case of mismatches, the maximum number of allowable mismatches is one. In the pilot study (**Chapter 3**), I allowed for 2 mismatches in order to maximize the number of miRNAs detected, albeit with lower confidence. For this set of libraries however, the libraries were large and of high enough quality for this to not be necessary – sequencing errors are unlikely enough not to require more than one mismatch.

5.3.2.2 Alignment

Using these parameters, I then ran the miRDeep2 mapper module (mapper.pl), aligning each of the 33 libraries to the mouse genome (GRCm38). The resultant (proprietary and required for downstream analysis) .arf format files are the equivalent of SAM format, giving details about the location on

the reference genome, mismatches (edit distance) and strandedness of each alignment.

Given the inevitable inclusion of parasite RNA in my libraries, I next attempted to identify putative *T. gondii* novel miRNAs from within my samples (**5.3.4**). Though of course any putative miRNAs of *T. gondii* origin would only be detectable in my infected samples, I thought it prudent to subject the uninfected samples to the same pipeline of discovery analysis as well, as a control. Thus, any novel miRNAs appearing only in the infected samples would have a greater confidence attached. For the same reason, though I used two different strains in my infections, I aligned all the libraries to both GT1 (version 13.0) and ME49 (version 13.0). GT1 is a Type I strain like RH, which was the strain used in the infections. Alignment results are shown in Figures 5.19 and 5.20.

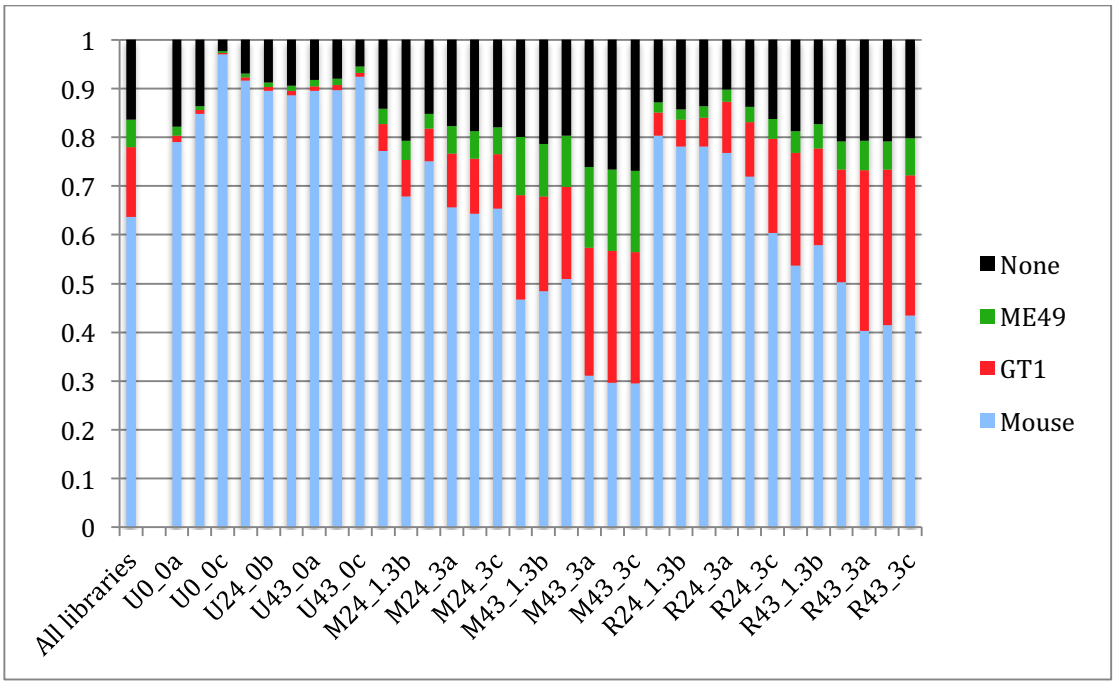


Figure 5.19. Alignment statistics to *M. musculus*, ME49 and GT1 – Raw trimmed reads

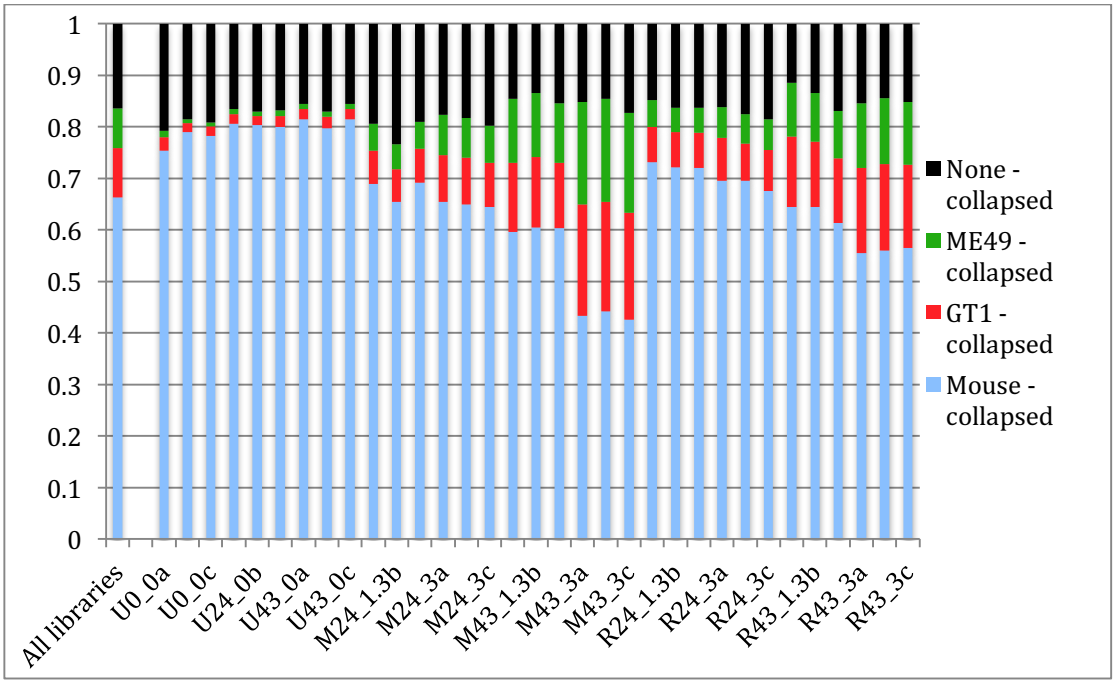


Figure 5.20. Alignment statistics to *Mus musculus*, ME49 and GT1 – Collapsed reads

In all three alignment scenarios, the proportion of reads that mapped to the parasite roughly follow the expectation of the sample's infection. That is to say, the highest proportion of ME49-mapping reads corresponds to libraries 16-18, made from cells infected with ME49 at an MOI of 3, for 43h.

The core miRDeep2 module consists of its 'quantifier' tool. This tool maps the known miRBase miRNAs to the precursor candidates and then intersects that with the mappings yielded by the sequencing reads.
After the quantification has been performed, there is a possible bifurcation in next steps: analysis of known miRNA expression (**5.3.5** onwards) or identification of putative novel miRNAs. (**5.3.3** and **5.3.4**).

5.3.3 Potential novel microRNAs from *Mus musculus*

At the highest scoring miRDeep2 tranche, 216 novel miRNAs were predicted. Of these, 18 had a seed sequence (2-8nt from the 5' end) identical to an miRBase-annotated mature miRNA from *Rattus norvegicus*, which was

inputted as the closest related species in miRBase. One of the 216 triggered an 'rfam alert' and was thus excluded. The full remaining list of novel putative *M. musculus* miRNAs in included in Appendix.

I then identified putative novel miRNAs that were differentially expressed over time, per condition, using the R package maSigPro (160). Subsequent to the time course analysis, I used MR-micro-T (161, 162) to predict putative targets for the differentially-expressed miRNAs. Of the predicted targets, I took the top tranche (with scores greater than 0.95) and subjected them to functional analysis using Reactome data from Mousemine (163).

Uninfected Sample – genes differentially expressed over time

Three miRNAs were significantly differentially-expressed across the uninfected time course, with profiles shown in the plots below. Tables of enriched pathways from predicted targets are shown below the expression profile figures, where available.

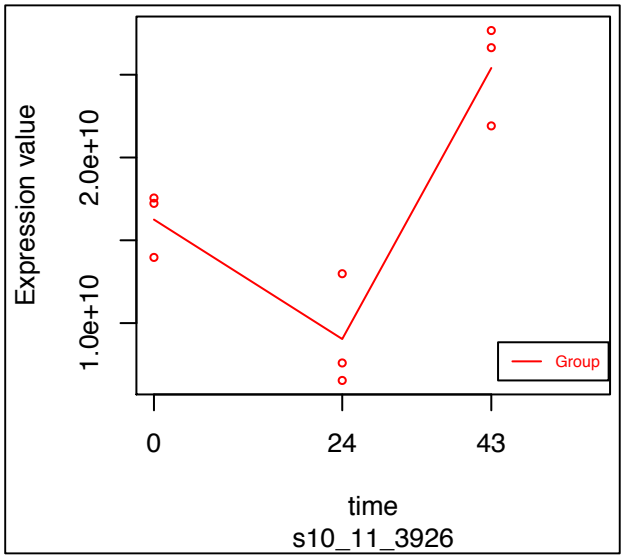


Figure 5.21. Expression Profile of putative novel miRNA s10_11_3926 – Uninfected

Table 5.2 Pathway enrichment of target genes for putative novel miRNA s10_11_3926

Reactome Pathway	adj p-value	No. of genes
Transcriptional regulation of white adipocyte differentiation	0.000351821	4
Transcriptional Regulation of Adipocyte Differentiation in 3T3-L1 Pre-adipocytes	0.000652876	4
Mus musculus biological processes	0.024080493	4

A total of 26 genes were predicted as putative targets for this miRNA. The four genes involved in all of the reactome pathways in Table 5.2 are: *Ncoa1*, *Med19*, *Med27* and *Ppargc1a*.

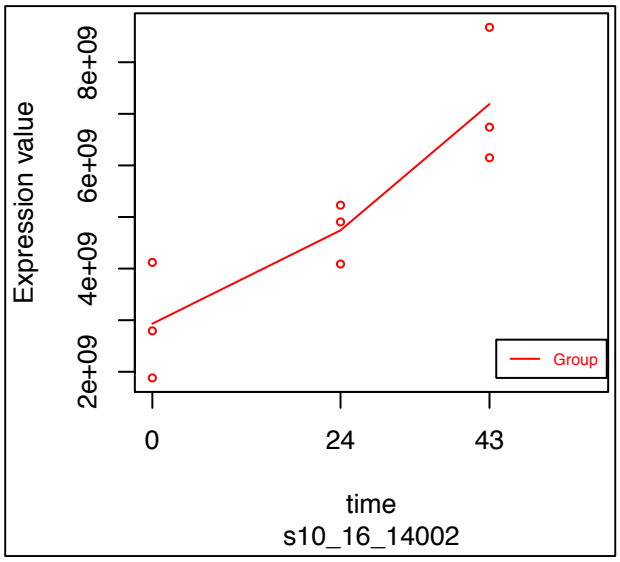


Figure 5.22. Expression Profile of putative novel miRNA s10_16_14002 – Uninfected

No statistically-significant pathway enrichment was found for the predicted targets of miRNA s_10_16_14002. The targets themselves are *Ammecr1*, *Edaradd*, *Foxp2*, *Kcna2*, *Ptprt* (Protein Tyrosine Phosphatase, Receptor Type T), and *Sos1*.

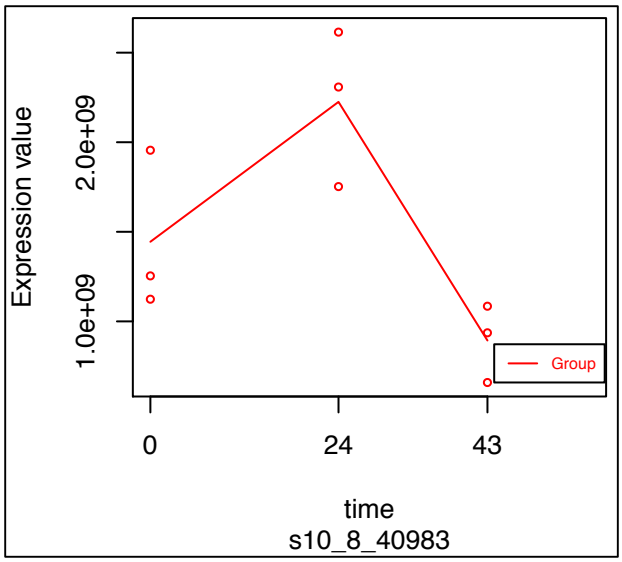


Figure 5.23. Expression Profile of putative novel miRNA s10_8_40983 – Uninfected

Again, no Reactome pathways were significantly-enriched for, but a GO term was enriched for: Organ growth. The genes that contributed to this process were: *Gja1* (gap junction protein, alpha 1), *Lats1* (large tumor suppressor), *Mtor* (mechanistic target of rapamycin (serine/threonine kinase)), *Smad2* (SMAD family member 2), *Tcf7l2* (transcription factor 7 like 2, T cell specific, HMG box), *Tgfbr3* (transforming growth factor, beta receptor III) and *Ttn* (titin).

ME49, MOI 1.2 – genes differentially expressed over time

Three miRNAs were found to be differentially-expressed across the time course. MicroRNA s_16_14002 was also differentially-expressed in the Uninfected sample, but with a very different profile across time. Given that the prediction of putative targets does not depend on the pattern of expression, the targets and pathways are the same as above.

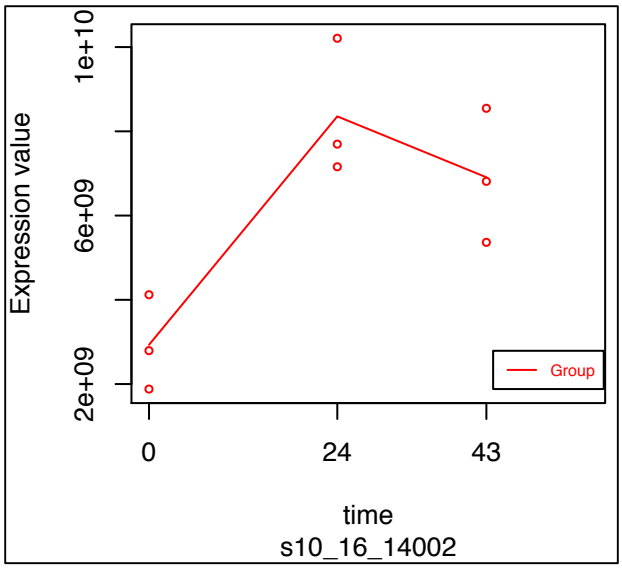


Figure 5.24. Expression Profile of putative novel miRNA s10_16_14002 – ME49, MOI 1.2

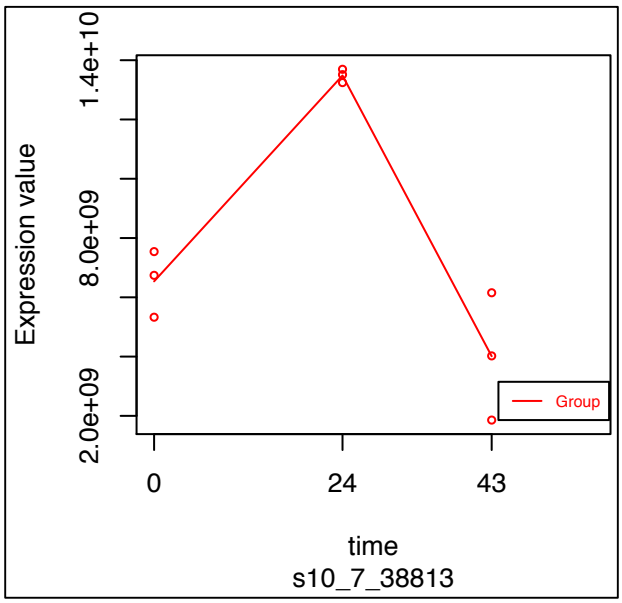


Figure 5.25. Expression Profile of putative novel miRNA s10_7_38813 – ME49, MOI 1.2

118 genes were predicted to be putative targets of miRNA S10_7_38813, but again, no Reactome pathways were enriched. That being said, the GO term analysis yielded several enriched terms, many having to do with metabolism and biosynthetic processes, Table 5.3.

Table 5.3 GO-term enrichment of target genes for putative novel miRNA s10_7_38813

GO Term	adj. p-value	#genes
regulation of nucleobase-containing compound metabolic process	0.001269326	41
regulation of macromolecule biosynthetic process	0.001623246	41
heterocycle biosynthetic process	0.001789389	41
aromatic compound biosynthetic process	0.002052568	41
regulation of cellular macromolecule biosynthetic process	0.002244811	40
regulation of biosynthetic process	0.002964252	42
regulation of transcription, DNA-templated	0.003021357	37
regulation of nucleic acid-templated transcription	0.003463961	37
regulation of nitrogen compound metabolic process	0.003510913	42
regulation of RNA biosynthetic process	0.003675824	37
nucleobase-containing compound biosynthetic process	0.003817873	40
regulation of macromolecule metabolic process	0.004196736	51
organic cyclic compound biosynthetic process	0.00437189	41
transcription, DNA-templated	0.005541801	37
regulation of cellular biosynthetic process	0.005803785	41
nucleic acid-templated transcription	0.006323055	37
macromolecule biosynthetic process	0.00693568	45
RNA biosynthetic process	0.007090377	37
negative regulation of cellular macromolecule biosynthetic process	0.008963633	22
regulation of RNA metabolic process	0.009113422	37
regulation of metabolic process	0.013500464	52
regulation of primary metabolic process	0.019623559	49
negative regulation of macromolecule biosynthetic process	0.020027057	22
cellular nitrogen compound biosynthetic process	0.022624369	43
organic substance biosynthetic process	0.02485083	49
cellular macromolecule biosynthetic process	0.025721011	43
regulation of cellular metabolic process	0.025785181	49
cellular biosynthetic process	0.03765248	48
negative regulation of cellular biosynthetic process	0.037890822	22
regulation of gene expression	0.039640641	40
biosynthetic process	0.040949275	49

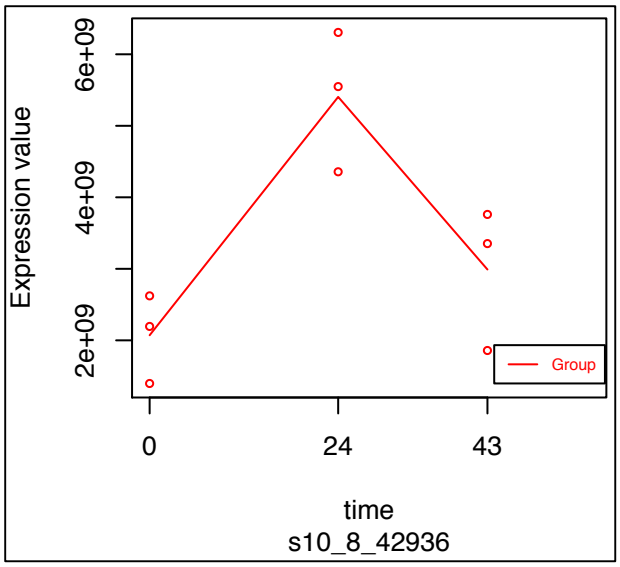


Figure 5.26. Expression Profile of putative novel miRNA s10_8_42936 – ME49, MOI 1.2

28 putative targets were identified, but neither GO Term nor Pathway enrichment yielded any statistically-significant results – this is likely due to the preponderance of "predicted genes" within the putative targets (over half).

ME49, MOI 3 – genes differentially expressed over time

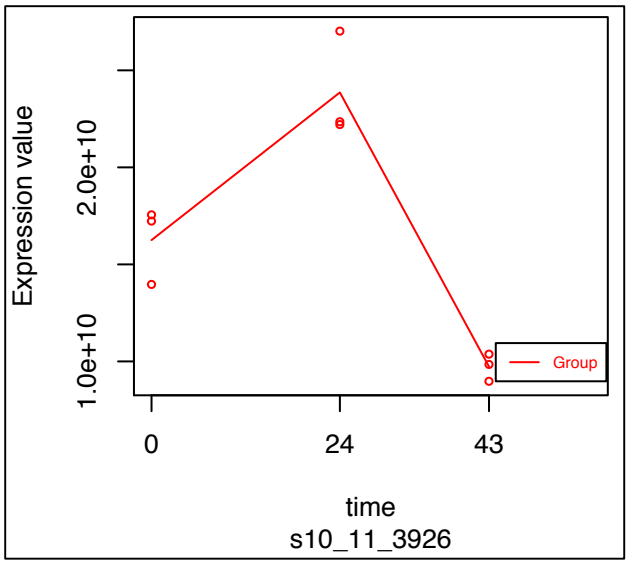


Figure 5.27. Expression Profile of putative novel miRNA s10_11_3926 – ME49, MOI 3

As with predicted miRNA S10_16_14002, S10_11_3926 was also significantly differentially expressed in the uninfected sample, though with a visibly different temporal profile.

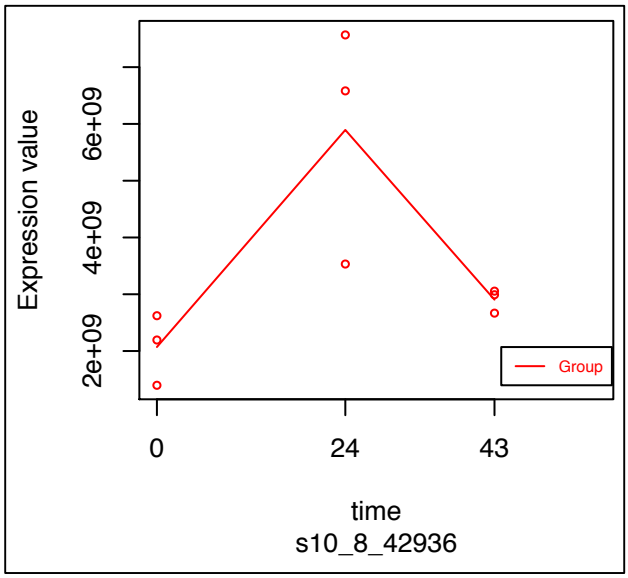


Figure 5.28. Expression Profile of putative novel miRNA s10_8_42936 – ME49, MOI 3.

5.3.4 Potential novel miRNAs from *Toxoplasma gondii*

Neither *T. gondii* nor any closely relate species are included in miRBase (even the previously identified miRNAs from other studies). That being said, miRDeep2 recommends the inclusion of even distantly-related species (113), so I opted to compare the two approaches: once without comparisons, and once using miRBase miRNAs from *Ectocarpus siliculosus*, a filamental brown alga.

In any case, given the absence of existing *T. gondii* miRNAs in miRBase for comparison, the scoring parameters that emerge from the miRDeep2 pipeline are slightly different than they were for the *M. musculus* samples. For instance, the estimated true/false positive score is based on the algorithm's performance in detecting known miRNAs within a particular set of samples and then using that metric to inform the likelihood of a putative miRNA being a true or false positive. Without reference miRNAs, this calculation is omitted from the analysis.

Though I ultimately performed the *T. gondii* miRNA discovery analysis using only infected samples, I re-ran the analysis using only the uninfected libraries – to ensure that no spurious miRNAs arising from mouse-derived reads were included in the novel set.

ME49

At a cut-off score of ten, six putative novel miRNAs were identified. Of these, two did not map to the genome itself, rather to unassembled supercontigs. Given that the lack of genome positioning was a criticism levelled at other attempts at miRNA identification in *T. gondii* (108), I opted to exclude these from my set. A further putative miRNA (provisional id 13393) was found to be predicted from uninfected reads alone, so it too was excluded, resulting in a remaining total of four novel putative miRNAs from ME49.

Table 5.4. Putative novel miRNAs from ME49

Putative miRNA 13393 (stricken out in the table) was only found in uninfected samples, indicating that is unlikely to be a true miRNA from *T. gondii*

T. gondii ME49 provisional id	Score with ***E. siliculosis***	Score w/out ***E. siliculosis***	***T. gondii*** ME49 Co-ordinates
13523	433.5	434.1	VIII:1266109..1266167+
~~13393~~	~~48.1~~	~~48.7~~	~~VIII:702568..702626+~~
18892	17	17.6	X:2773287..2773347+
18365	15.1	15.7	IX:5990908..5990990-
19667	9.7	10.3	X:6540009..6540050+

The inclusion of *E. siliculosis* neither improved the scores of putative miRNAs nor did it identify any new putative miRNAs with a score above 10. In fact, the only effect that this inclusion had was to identify one novel putative miRNA with the same seed sequence as an existing *E. siliculosis*

miRNA, but this putative *T. gondii* ME49 miRNA had a score of only three so was excluded from further analysis.

The relative sequence contribution of each sample to each putative miRNA is given below, as well as graphical representations of frequency density plots of the alignment stacks and the predicted precursor RNA secondary structure. NB. The pie charts are for collapsed reads. That is to say, if a particular library contributed x reads to the alignment, but that read had been sequenced multiple times, that sample's contribution would still be represented as x in the pie chart.

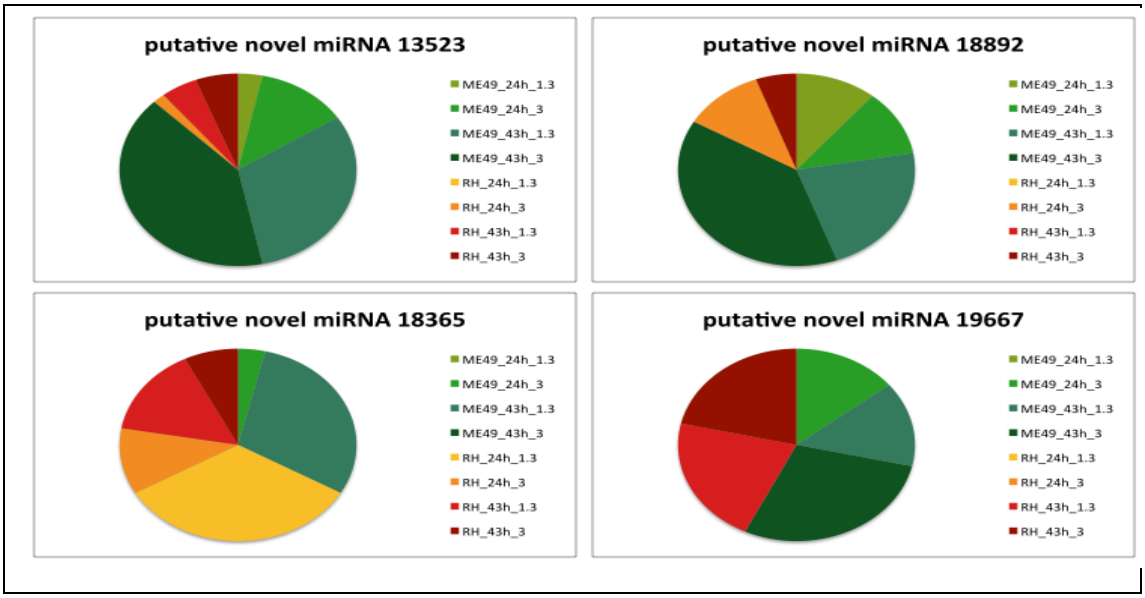


Figure 5.29. Relative sequence contribution of each sample to the putative novel ME49 miRNAs. Collapsed counts from each experimental samples were assessed for their proportional contribution to the putative novel miRNA. In all but one case, the preponderance of reads were contributed by ME49.

GT1

At a cut-off score of ten, four putative novel miRNAs were identified, all of which could be positioned on the GT1 genome.

Table 5.5. Putative novel miRNAs from RH. Putative miRNAs in green align exactly to putative miRNAs discovered in ME49, while putatuve miRNA 20653 was excluded from further anaysis since it's score without including *E. siliculosis* was under ten.

id	Score with *E. siliculosis*	Score without *E. siliculosis*	*T. gondii GT1* co-ordinates
14134	429.5	430.1	chrVIII:1260791..1260849+
28695	45.2	45.8	chrXII:6281695..6281774-
28696	36.8	37.4	chrXII:6281758..6281799-
19812	17	17.6	chrX:2684478..2684538+
~~20653~~	~~10.3~~	~~9.7~~	~~chrX:6429927..6429968+~~

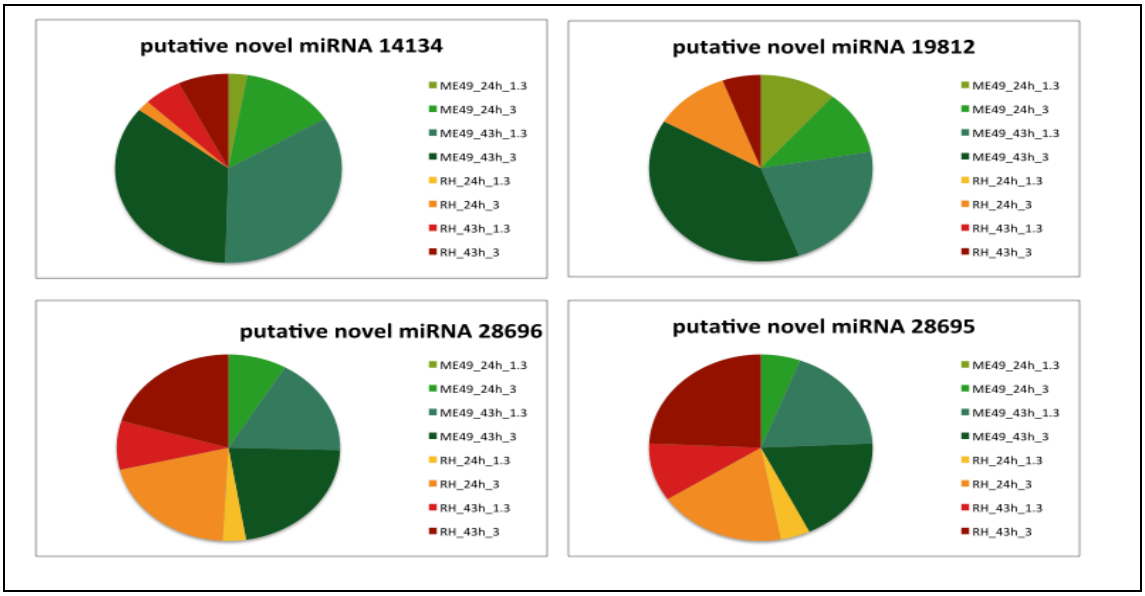


Figure 5.30. Relative sequence contribution of each sample to the putative novel ME49 miRNAs. Collapsed counts from each experimental samples were assessed for their contribution to the putative novel miRNA.

Given the possibility – however remote – of parasite miRNAs targeting mouse mRNAs directly, I used the algorithm above (MR-MicroT (161, 162) to predict putative *M. musculus* targets.

Predicted *M. musculus* targets of putative ME49 miRNAs

Provisional id 13523:

No putative targets were predicted above a score of 0.95.

Provisional id 18892:

Of the nine putative targets predicted to be targeted by this putative miRNA, three were unestablished genes (e.g. predicted, or RIKEN cDNA probes). The remaining six were: *Arl4c* (*ADP-ribosylation factor-like 4C*), Cnot2 (CCR4-NOT transcription complex, subunit 2), Erp44 (*endoplasmic reticulum protein 44*), *Peak1* (*pseudopodium-enriched atypical kinase 1*), *Slc35f1* (*solute carrier family 35, member F1*) and *Zfp938* (*zinc finger protein 938*).

Provisional id 18365:

Given the low-complexity nature of this miRNA (its sequence being UGUGUGUGUAUGUGUGUAUGUG), specificity in the prediction of targets is extremely difficult. As such, the list of putative targets is very large (298 at a score of >=0.95).

Provisional id 19667:

A single gene was predicted to be targeted by this putative miRNA, *Nsrp1* (*nuclear speckle regulatory protein 1*).

Predicted *M. musculus* targets of GT1 putative miRNAs

Of the putative miRNAs identified by alignment to GT1, only two were not also found in the ME49 set.

Provisional id 28695:

A single predicted target was identified, *Cadm2* (*cell adhesion molecule 2*).

Provisional id 28696:

Two putative target genes were identified, *Map2kp* (*mitogen-activated protein kinase kinase 4*) and Il1rapl1 (*interleukin 1 receptor accessory protein-like 1*)

5.3.5 Differentially Expressed miRNAs

While the miRDeep2 'quantifier' module does yield count data for known miRNAs, it lacks the capacity for either normalization or statistically-sound differential expression analysis – it merely returns naïve/raw count data for each miRNA and each library. Given the vastly better quality of my libraries (as compared to the pilot study in **Chapters 3** and **4**), I opted to use EdgeR for both these analysis steps. Before normalising the data, EdgeR offers (and recommends) two filtering steps, the first being to filter out tags that have 'zero' value across all conditions. A further filtering step also removes tags that are represented by fewer than one CPM (count-per-million) in fewer than two samples. This step ensures that only likely-expressed candidates are considered in downstream analyses, thus reducing the risks of multiple testing. Filtering on CPM, rather than raw read counts also provides some measure of scaling (given that library sizes can vary a great deal without much relation to the underlying biological reality). EdgeR utilises TMM normalisation and negative binomial distributions especially geared toward count data (rather than, for instance, microarray data).

A total of 274 microRNAs were found to be differentially expressed with regard to the uninfected sample, at any time following infection. How these were spread across strain, MOI and time point are presented as Venn diagrams in Figure 5.31.

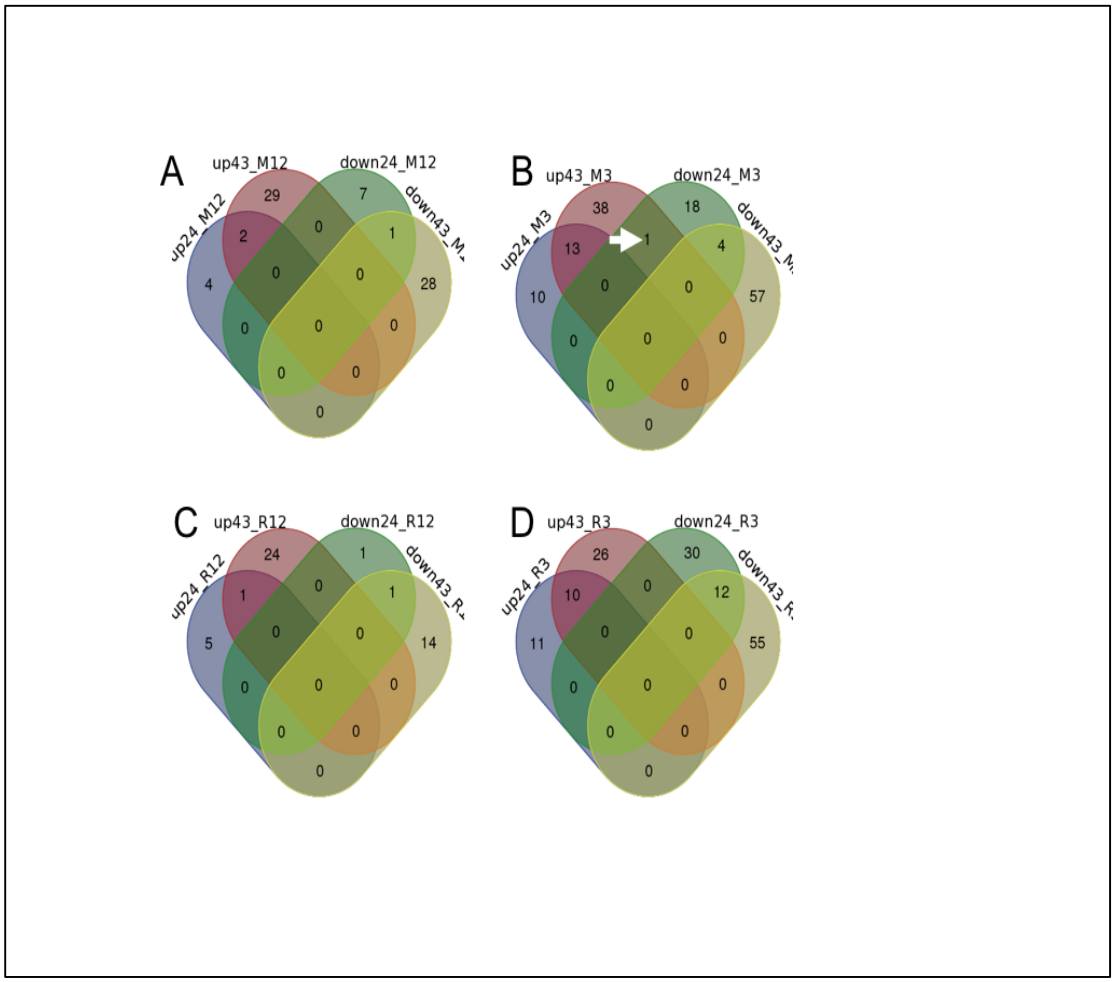


Figure 5.31. Intersection of host miRNAs identified as differentially expressed among all conditions.

Only one miRNA (olive green with a white arrow in Figure 5.31) was found to be differentially expressed in 'opposite directions' over time (that is to say, underexpressed at 24h and then overexpressed at 43h, or vice versa); mmu-miR-425-3p.

I then examined the temporal pattern of miRNA differential expression (NB. I excluded mmu-miR-425-3p from these charts due to its atypical[8] biphasic pattern of expression upon ME49 MOI 3 infection). There are two different methods of examining this issue of temporal expression:

1) Expression that is induced (or suppressed) at 24h and then may (or may not) be sustained through 43h
2) Expression that is induced (or suppressed) *only* at 24h or 43h.

Perhaps the more interesting case, is 2), which is what is represented in Figures 5.32 and 5.33. This could reveal where 'waves' of transcriptional initiation might be observed, i.e. looking at miRNAs whose induction/suppression was *first observed* at a particular time point (so, for instance, a gene that was induced at 24h and then sustained through 43h would be excluded from the 43h group). It would be too much to say that host miRNAs induced (or suppressed) by *T. gondii* infection follow a biphasic mode of expression, given that only two time points past zero were sampled, but this nevertheless suggests that they are not dysregulated 'wholesale', at the same time point.

[8] Atypical with regard to the rest of the dysregulated miRNAs in my dataset.

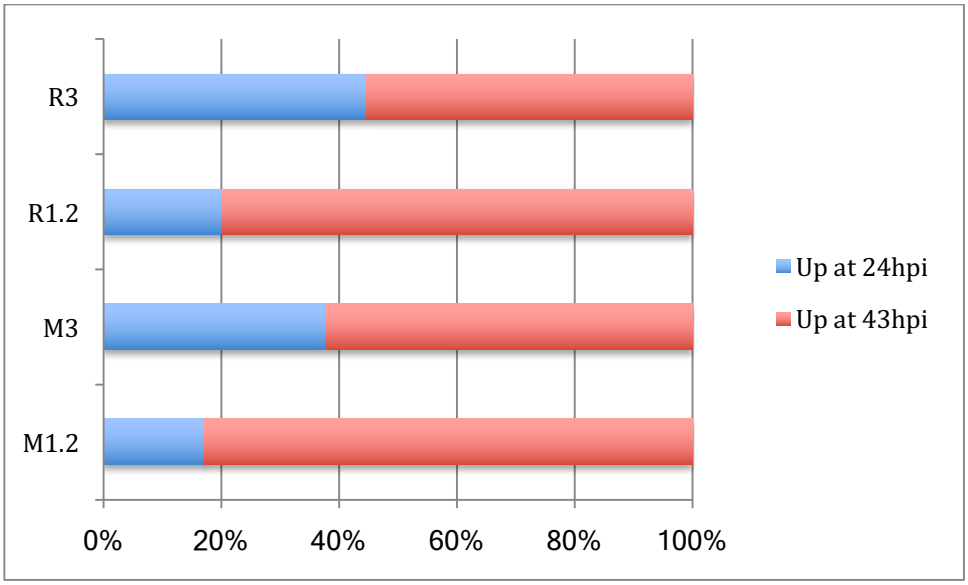


Figure 5.32. Timing of miRNA upregulation. microRNAs were scored and grouped based on the timepoint at which they first emerged as upregulated. The microRNA genes appear to follow 'waves' of transcriptional induction.

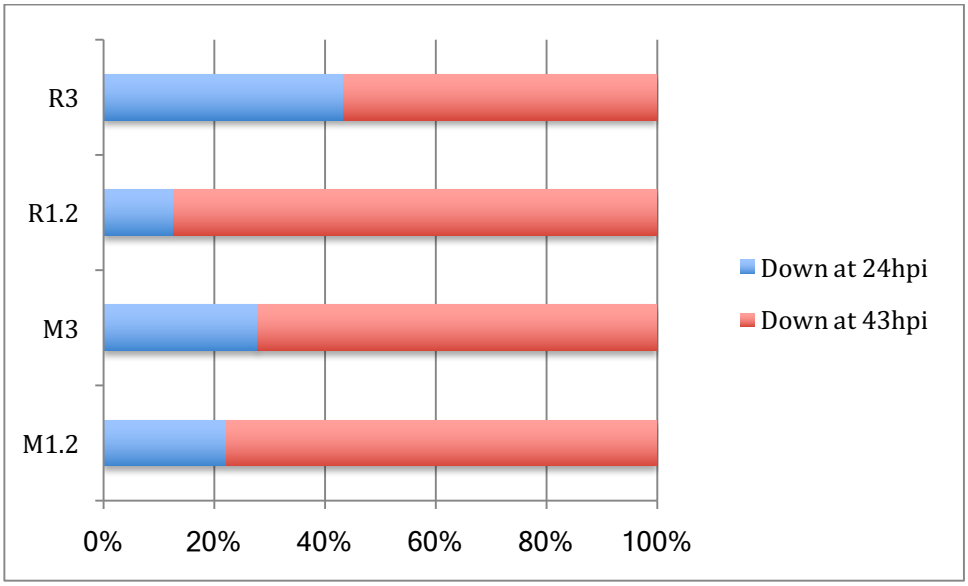


Figure 5.33. Timing of miRNA downregulation. microRNAs were scored and grouped based on the timepoint at which they first emerged as downregulated. The microRNA genes appear to follow 'waves' of transcriptional induction.

Given that (in both strains) the lower MOI showed a greater proportion of dysergulated miRNAs at the later time point, I thought that perhaps this might be due to parasite burden. There is an argument to be

made that a higher MOI might simply represent a higher (or lower) fold-change in the same miRNA but with an 'early onset'. To test this hypothesis, I looked at dysregulated miRNAs within each strain, comparing ones that were dysregulated beginning at 24hpi in the higher MOI with those that were dysregulated at 43hpi (and no earlier) in the lower MOI. This hypothesis would imply that the higher MOI would be 'one step ahead' in terms of dysregulation, that is, that genes upregulated at the higher MOI at 24hpi would be followed by the same ones in the lower MOI, only later.

RH – Delayed Onset

For the Type I strain, of the 21 genes that were upregulated at 24h at MOI 3, five (24 per cent) exhibited a statistical 'delayed' upregulation (i.e. were overexpressed at 43 h and not at 24h). In the downregulated set of miRNAs, only 2 of the 22 genes downregulated at 24hpi were 'followed' by a similar downregulation at 43hpi in the lower MOI.

ME49 – Delayed Onset

In ME49, the 'delayed onset' hypothesis is strengthened, with over half of the 24h, MOI 3 overexpressed genes being expressed later (and only later) at the lower MOI. Only 5 of the genes that were downregulated at MOI 3 24hpi (out of 42 in total) were downregulated only at 43hpi at MOI 1.2.

5.3.6 Common Core of Dysregulated miRNAs

Given the fact that the generalised host miRNA response to *T. gondii* infection has still not been well-characterised, I decided to then concentrate on those dysregulated miRNAs that were common to *T. gondii* infection as a whole, rather than related to strain-specificities. To do this, I intersected the genes that were upregulated at any time in both strains, at both MOIs,

resulting in 38 upregulated miRNAs and 55 downregulated ones (Tables 5.6 and 5.7).

Table 5.6. *Mus musculus* microRNAs upregulated in infection, by both strains

mmu-miR-7081-3p	mmu-miR-1934-5p	mmu-miR-98-5p	mmu-miR-146a-5p
mmu-miR-877-5p	mmu-miR-5121	mmu-miR-7650-5p	mmu-miR-5103
mmu-miR-7655-5p	mmu-miR-6240	mmu-miR-677-5p	mmu-miR-98-3p
mmu-miR-582-5p	mmu-miR-582-3p	mmu-miR-1931	mmu-miR-132-3p
mmu-miR-9-5p	mmu-miR-1947-5p	mmu-miR-295-3p	mmu-miR-149-5p
mmu-miR-101a-5p	mmu-miR-1194	mmu-miR-217-5p	mmu-miR-9-3p
mmu-miR-1948-5p	mmu-miR-708-5p	mmu-miR-5099	mmu-miR-155-3p
mmu-miR-708-3p	mmu-miR-505-5p	mmu-miR-188-5p	mmu-miR-1191a
mmu-let-7j	mmu-miR-7042-5p	mmu-miR-155-5p	
mmu-miR-3086-5p	mmu-miR-1948-3p	mmu-miR-1945	

Table 5.7. *Mus musculus* microRNAs downregulated in infection, by both strains

mmu-miR-690	mmu-miR-299b-3p	mmu-miR-671-3p	mmu-miR-467b-5p
mmu-miR-30b-3p	mmu-let-7b-3p	mmu-miR-299a-3p	mmu-miR-503-5p
mmu-miR-140-5p	mmu-miR-466p-5p	mmu-miR-92b-3p	mmu-miR-199a-3p
mmu-miR-574-5p	mmu-miR-30c-1-3p	mmu-miR-299b-5p	mmu-miR-3072-5p
mmu-miR-193b-3p	mmu-miR-199a-5p	mmu-miR-671-5p	mmu-miR-382-5p
mmu-miR-450b-3p	mmu-miR-466k	mmu-miR-328-5p	mmu-miR-485-3p
mmu-miR-5126	mmu-miR-134-5p	mmu-miR-154-3p	mmu-miR-467a-5p
mmu-miR-6238	mmu-miR-410-5p	mmu-miR-351-5p	mmu-miR-411-3p
mmu-miR-377-3p	mmu-miR-187-3p	mmu-miR-3068-5p	mmu-miR-154-5p
mmu-miR-135b-3p	mmu-miR-134-3p	mmu-miR-125a-3p	mmu-miR-26a-2-3p
mmu-miR-380-5p	mmu-miR-379-3p	mmu-miR-299a-5p	mmu-miR-615-5p
mmu-miR-199b-3p	mmu-miR-129-5p	mmu-miR-379-5p	mmu-miR-1247-5p
mmu-miR-666-5p	mmu-miR-129-2-3p	mmu-miR-193a-5p	mmu-miR-574-3p
mmu-miR-127-3p	mmu-miR-466i-5p	mmu-miR-466h-3p	

5.3.7 Functional Analysis of the Dysregulated miRNAs

While a number of well-characterised miRNAs appear in both lists (miR-155, miR-125, miR-146 and miR-92 for instance) the relatively large number of dysregulated miRNAs led me to perform a functional enrichment analysis bioinformatically, using WebGestalt (156, 157). I initially opted for a 0.05 significance level (with Benjamini-Hochberg multiple testing correction) but relaxed this to 0.1 for the downregulated miRNAs, given that no significant results were found that this level.

5.3.7.1 Downregulated miRNAs, Upregulated Targets

The only pathway to be significantly enriched (even at a significance level of 0.1) by this list of target genes was Sphingolypid Metabolism.

5.3.7.2 Upregulated miRNAs, Downregulated Targets

More pathways were significantly enriched using the set of targets from the downregulated miRNAs, they are shown in Table 5.8.

Table 5.8 Enriched KEGG pathways for predicted targets of miRNAs upregulated by the common core of dysregulated miRNAs

Osteoclast differentiation
Toll-like receptor signaling pathway
Jak-STAT signaling pathway
Neurotrophin signaling pathway
Hepatitis C
Chemokine signaling pathway

5.3.8 Profiles of Selected miRNAs

Of the two miRNAs that were identified as being dysregulated in **Chapter 4** (mmu-miR-200b-3p and mmu-miR-3080-5p) only the former was present in my more comprehensive dataset here (not just the common core, but anywhere). To check whether the directionality of differential expression was

common between the two analyses, I plotted profiles of mmu-miR-200b-3p for each strain (**Figure 5.34**). Note that the Chapter 4 analyses were performed on NIH/3T3 infected with ME49 at an MOI of 5 for 24h.

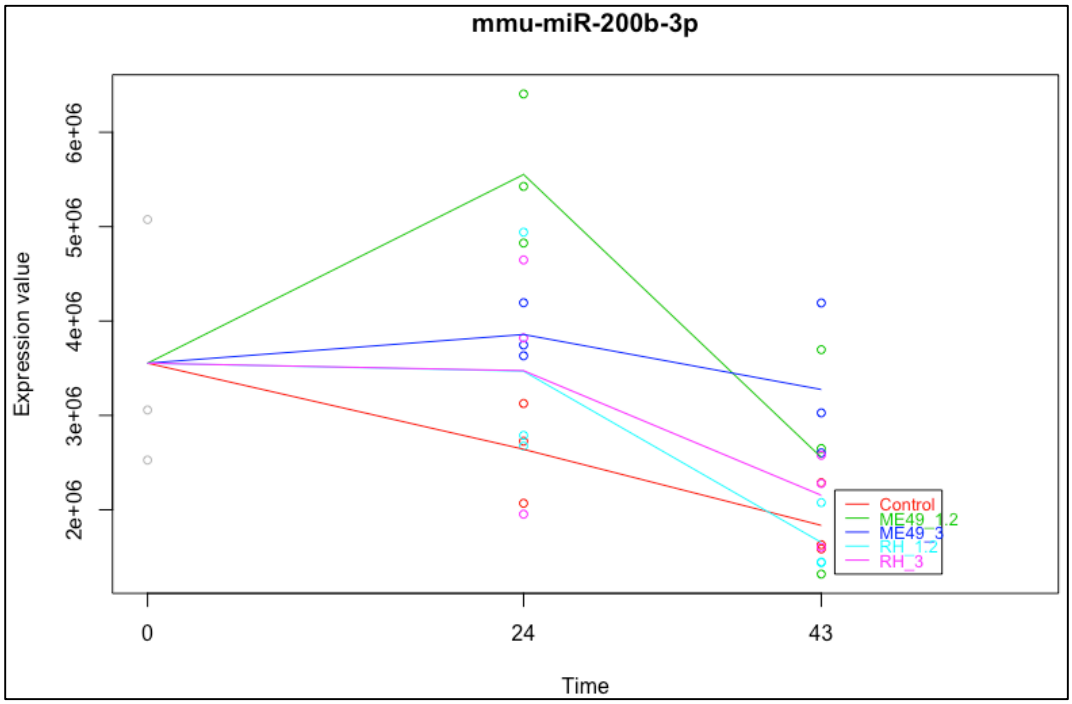


Figure 5.34. Profiles of mmu-miR-200b-3p

Other noteworthy miRNAs are profiled below in **Figures 5.35** to **5.38**).

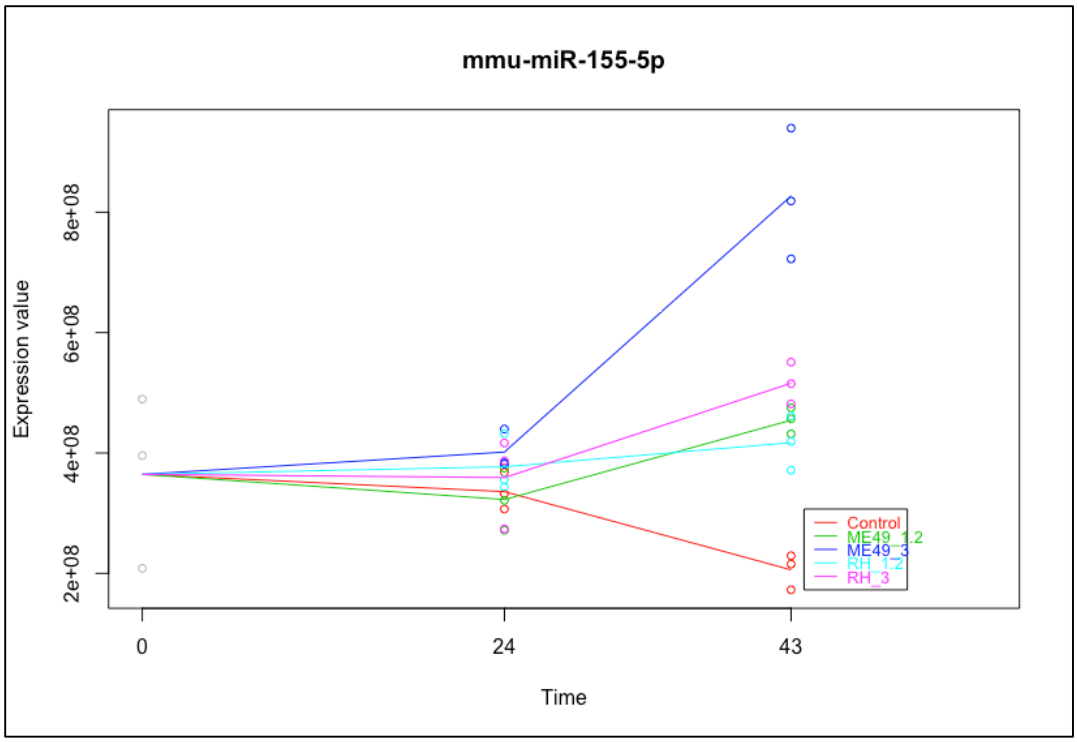


Figure 5.35. Profiles of mmu-miR-155-5p

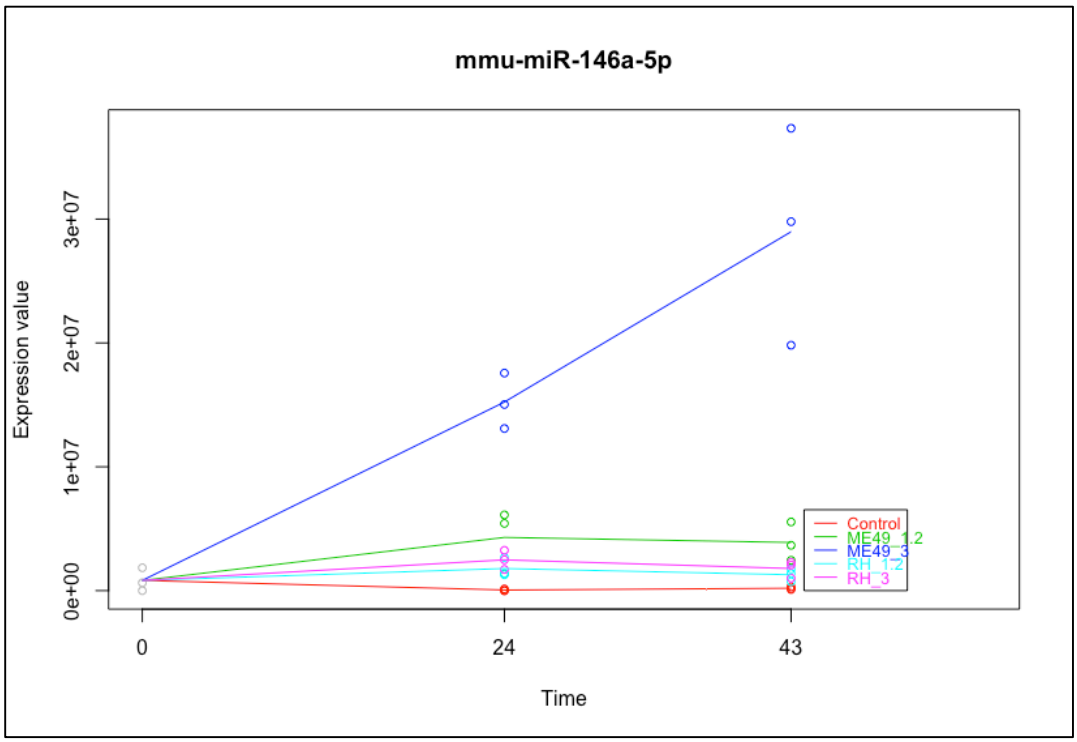


Figure 5.36. Profiles of mmu-miR-146a-5p

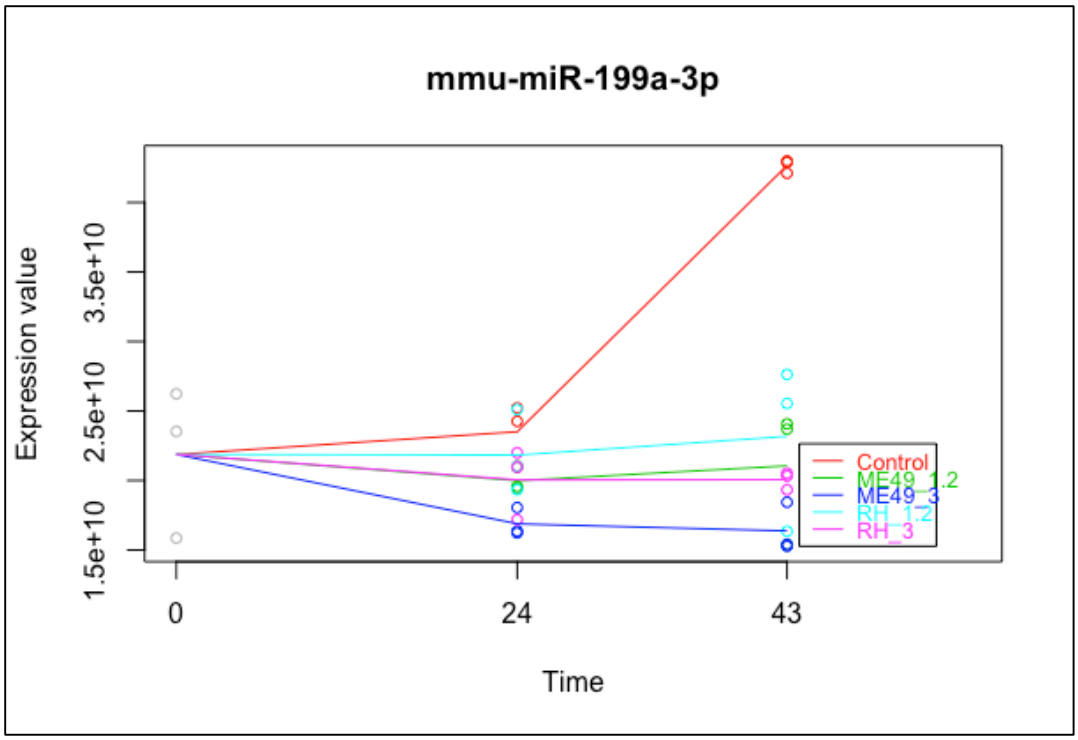


Figure 5.37. Profiles of mmu-miR-199a-3p

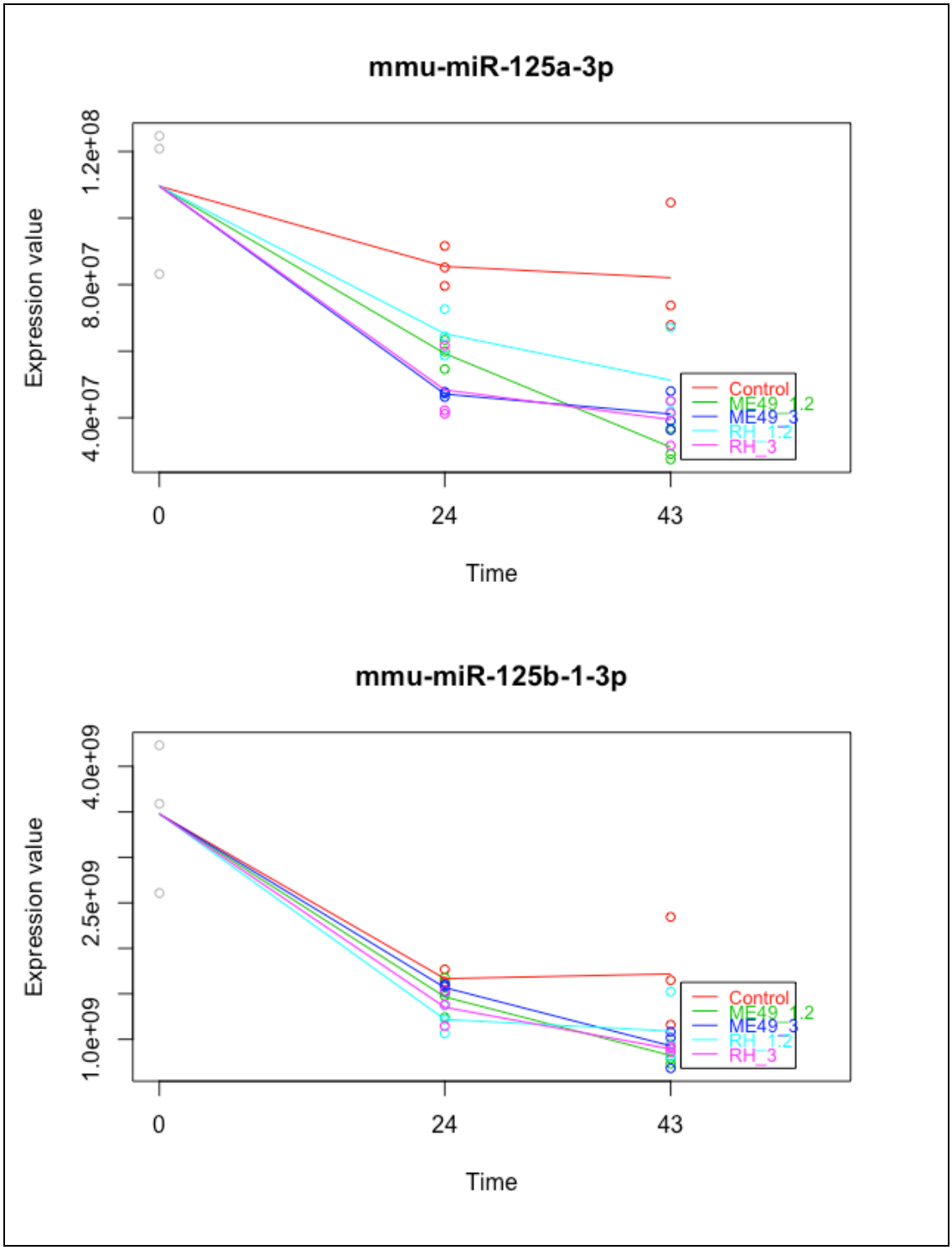
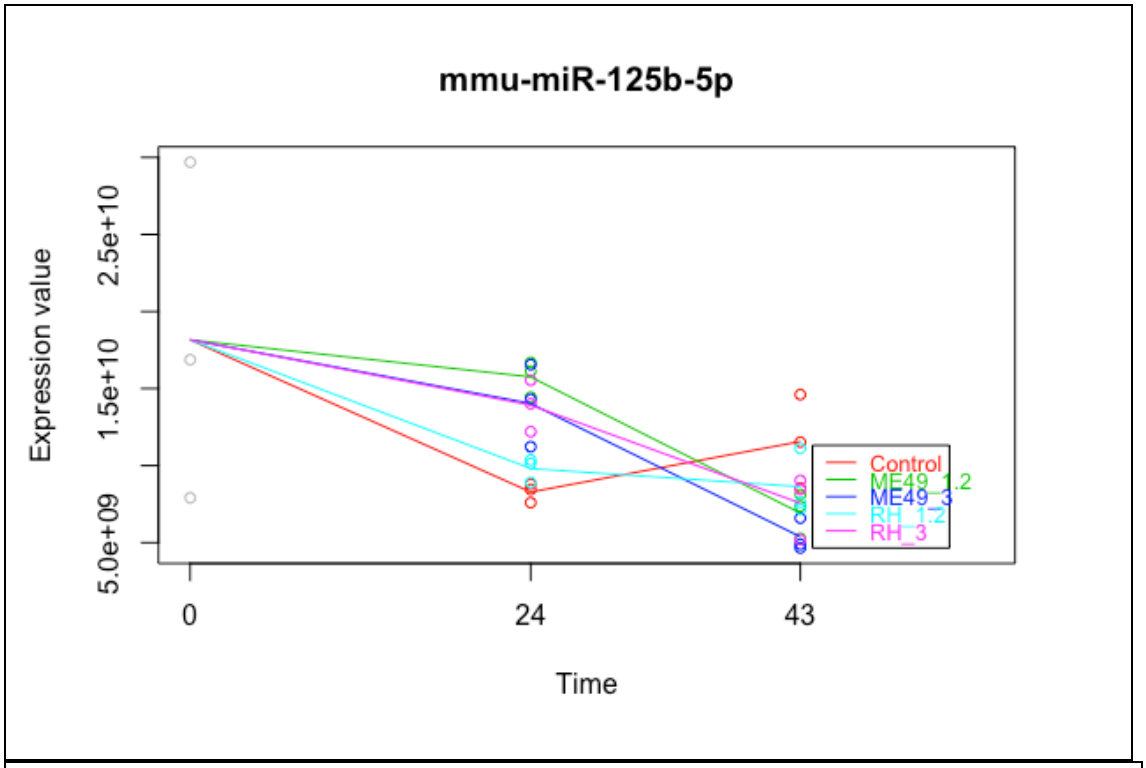


Figure 5.38. Expression profiles of miR-125 family members



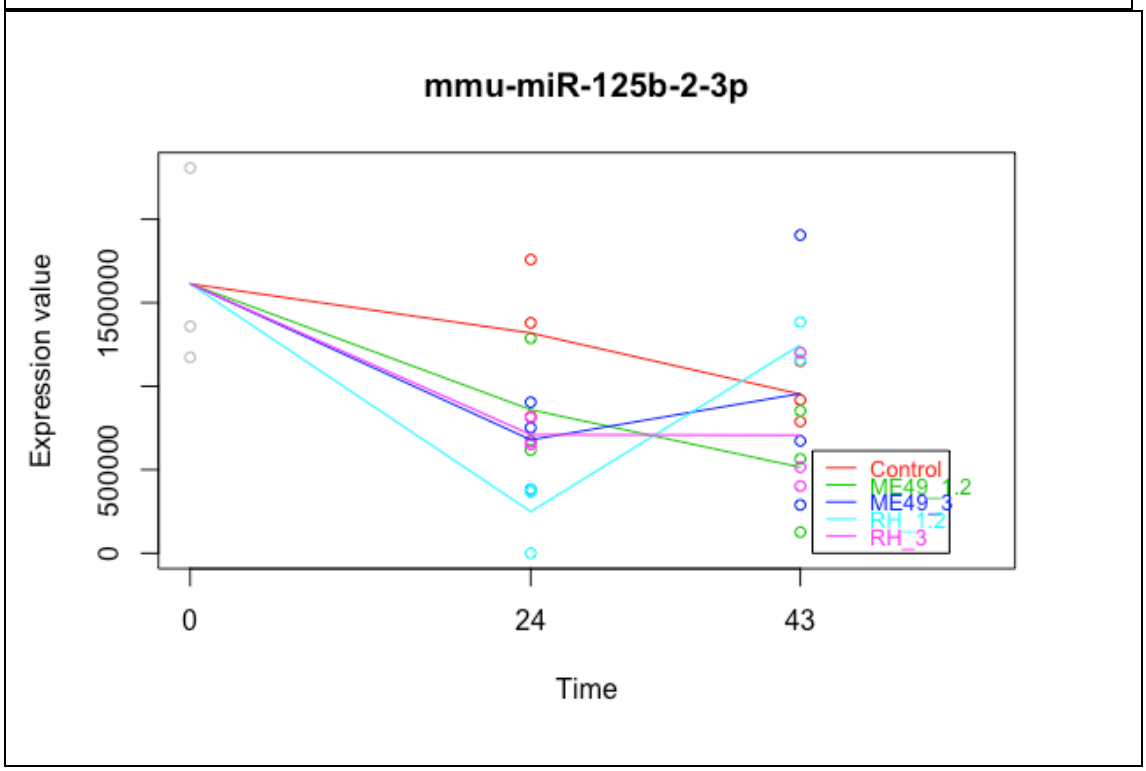


Figure 5.38, continued. Expression profiles of miR-125 family members

Given the known upregulation of MYC in *T. gondii* infection (21) and that MYC has been shown to directly and specifically suppress miR-23a/b (164), I also looked at those two miRNAs (**Figure 5.39**).

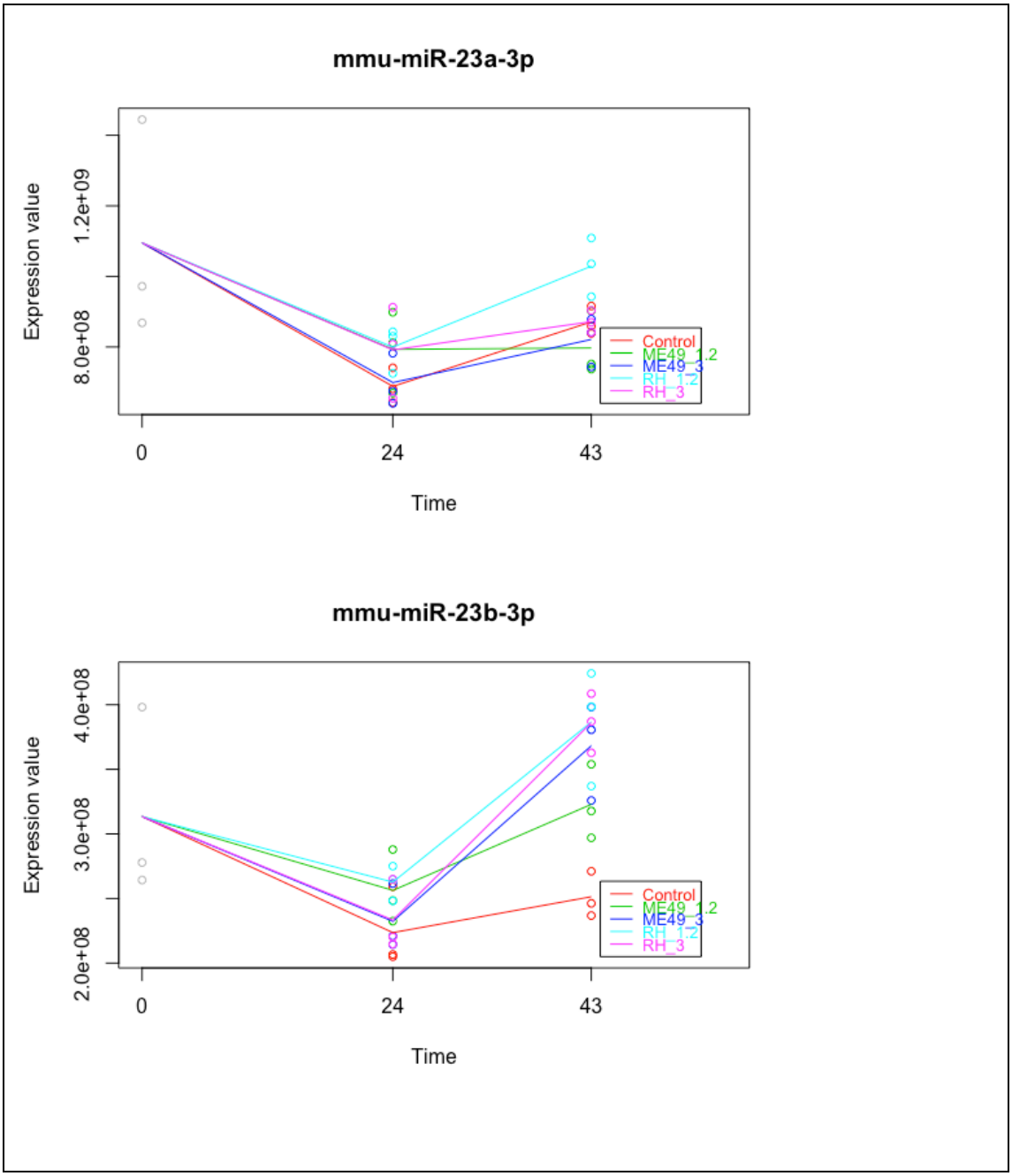


Figure 5.39. Expression profiles of mmu-miR-23a-3p and mmu-miR-23b-3p

5.4 Discussion

My analyses report a number of host cell miRNAs dysregulated due to infection by *Toxoplasma gondii*, as well as a number of novel miRNAs, both host- and parasite-derived.

5.4.1 Targets of miRNAs

The nature of miRNA analysis – using miRNA targets as input for functional analyses by pathway enrichment – is inherently dependent on the quality of the targets. This becomes particularly crucial when looking at large lists of miRNAs that do not lend themselves to individual analysis.

Though forward genetics have not been widely used to identify miRNA genes, their obvious advantage is that, at the very least, a scoreable pheonotype exists that can then be used to infer the function of the miRNA. Instead, we now find ourselves in the position where hundreds of miRNAs have been identified (and validated) in numerous organisms, but with relatively few functional annotations. Until relatively recently, most of the work in this area has concentrated on the computational prediction of miRNA targets. A miRNA's characterstic mode of action relies on the formation of a duplex between the miRNA and the 3'UTR of its target. Unlike in plants, base-pairing between the miRNA and the target is usually imperfect, which obviously makes target prediction much more difficult.

Since 2009, the process of target prediction and validation has been made far more reliable by the use of HITS-CLIP (165) (indeed, of the 39,534 targets that I used as my background list for the WebGestalt analysis, over 73% were validated using HITS-CLIP). HITS-CLIP works through UV crosslinking to purify RNA-protein complexes. Chi et al used this method with Argonaute protein, which binds to both the miRNA and its mRNA targets – which can then be precisely identified.

MR-MicroT (161, 162) is one of very few target prediction algorithms that allows the search of novel putative miRNAs' targets (and even this programme is still in beta). To my knowledge, no independent validation of the process has yet been performed, and so those results must be taken as highly preliminary. Quite apart from the fact that the potential targets of these putative miRNAs are themselves ill-defined, the use of pathway enrichment on them to identify the potential function of putative novel miRNAs is extremely poorly defined. After all, no reasonable background correction is possible and so, these results must be taken with great reservations.

5.4.2 Novel miRNAs – *Mus musculus*

As with most classic model organisms, mouse miRNAs have, since their discovery, been well-studied: of 35,827 miRNAs in miRBase 21, 1,915 are mouse miRNAs, second only to human miRNAs (2,588 of which are listed). As a result, my expectation for my set of 33 libraries was that, were any novel mouse miRNAs to be detected, they would most likely appear as mouse miRNAs dysregulated specifically in response to infection by *Toxoplasma gondii*. At first this appeared to have been somewhat borne out by the fact that none of the novel miRNAs were predicted using uninfected samples alone.

Overall, my analyses identified a total of 215 novel miRNAs. However, upon time course analysis, I then identified three novel miRNAs that were significantly differentially-expressed over the course of the 43h experiment, solely in uninfected NIH/3T3 cells. This indicates that these novel miRNAs could be dysregulated in response to confluence and contact inhibition. Indeed, of the three, though only one had predicted targets with significant pathway enrichment, these pathways corresponded to adipocyte differentiation. While NIH/3T3 are not known to spontaneously differentiate into adipogenic cells, given that most adipocyte research is conducted using

late-stage 3T3-L1 cell lines, it is not inconceivable that this particuar miRNA (identified in my analysis as s10_11_3926) has an impact on post-proliferation processes in 3T3 or 3T3-like cells. An important study of confluence and miRNA expression was conducted by Hwang et al (166), where a comparison was made of primary human fibroblast, NIH/3T3 and HeLa cell lines' microRNA expression. Overall, they found that, in all three cases, growing cells towards confluence resulted in a global increase in miRNA biogenesis, an effect that they attribute to increased cell-cell contact (rather than, say, quiescence). Hwang et al. conducted northern blots to assess the presence of a battery of miRNAs and, following qPCR analysis, concluded that the increased biogenesis of miRNAs upon confluence was not transcriptional in nature (except for one case, miR-34a). However, the qPCR experiments were not conducted on NIH/3T3 cells. Even if the majority of confluence-dysregulated miRNA biogenesis can be attributed to posttranscriptional processing, this does not preclude the possibility that a few specific miRNAs are transcriptionally dysregulated – including some that have not yet been discovered – when profiled across a long growth period, as my results indicate.

5.4.3 Novel miRNAs – *Toxoplasma gondii*

Far fewer novel miRNAs (or miRNA-like RNA species) were identified from the parasite. This is to be expected, purely from a technical standpoint, given the far lower alignment coverage. After all, parasite contribution to the RNA sample would have been small. The target gene predictions for these genes must be interpreted in an even more cautious fashion: while it has been hypothesised (112), there is no evidence at all that this might be the case.

5.4.4 Differentially Expressed Known Host miRNAs

My results suggest that the host miRNA response to *T. gondii* infection is in fact rather modest – 274 miRNAs were found to be differentially expressed compared to the uninfected sample, at any time and by any strain/MOI. That being said, some of the most well-characterised miRNAs did emerge.

There is some discussion as to whether analyses of differential expression are best performed considering lists of up/down-regulated genes/miRNAs separately or together. Single annotated networks (e.g. within KEGG) can contain genes – in this case miRNA target genes – that are both up and down-regulatory towards the pathway as a whole. That is to say, a particular KEGG pathway being enriched for may not necessarily indicate that that pathway is being driven "positively". Rather, this may only indicate that several members of that pathway (whether positive or negative regulators) were disproportionately represented in the sample list. While some attempt has been made to quantify this in 'simple' gene expression analysis (167), no such study has been performed that takes into account the extra ('inverted') layer of complexity added by miRNA control and so this may represent a source of added complexity.

That being said, a number of pathways that are known to be dysregulated by *T. gondii* infection did appear in the enrichment. The parasite is known to downregulate STAT1, though a miRNA component to that downregulation has not been explored. *Toxoplasma gondii*'s ability to interfere with the host cell cytokine pathways – many with NFKB implications - has also been examined but not in terms of miRNA regulation beyond the role of miR-146a, which I find in my dataset.

Given the somewhat tentative nature of this pathway analysis, I focus instead on the well-characterised miRNAs that emerged from my list.

Perhaps unsurprisingly, given its multiple and varied roles, the very 'famous' let-7 pathway makes several appearances in my dataset. What is on

first glance very curious, however, is that members of this family appear both in the up- and downregulated lists: mmu-miR-98 and mmu-let7j are upregulated while mmu-let7b is downregulated (Tables 5.6 and 5.7).

One of the difficulties in miRNA research is that, given the relative infancy of the field, very basic issues such as nomenclature and standardisation change rapidly and often. This means that early work on miRNA family members is often difficult to relate to current nomenclature: one can imagine a scenario[9] where, for instance, early literature ascribes a particular function to let-7 without specifying which family member was being considered (*M. musculus* alone contains 12 different let-7 family members). Absent sequence information, it is therefore difficult to interpret that kind of experiment and decide which family member was being referenced.

It is true also that members of a single miRNA family can have vastly different sets of target genes, and it is still unclear how these interactions are mediated, mechanistically. For the let-7 family, due at least in part to the nomenclature issue highlighted above, even despite it being the most well-studied miRNA family, the exact similarities and differences between the family members is still poorly-understood. Of the family members in my dataset, mmu-let-7b and mmu-miR-98 are far better studied than mmu-let-7j[10] so I will concentrate on them. Wang et al found in 2011 that miR-98 directly regulates Fas: miR-98 mimics decreased the levels of Fas-induced apoptosis in HeLa cells (and inhibitors increased it). Thus, its upregulation in infected cells is consistent with the observation that *T. gondii*-infected cells exhibit lower levels of apoptosis, including Fas-dependent (34).

Given that my dataset included the well-characterised mmu-miR-125a, I generated profiles of other family members (Figure 5.38). Among the family's roles is the regulation of NFkB. The potential mechanism that Kim et

[9] Actually, one does not even need to "imagine"

[10] A PubMed search for each revealed 126 and 299 results for "miR-98" and "let-7b" respectively, while "let-7j" yielded only a single paper – a survey of miRNAs in Nile tilapia.

all suggests is via targeting of *TNFAIP3* (168). That miR-125 is itself a target of NFkB adds another layer of complexity to an already complex regulatory mechanism. Another role for miR-125 is in the regulation of apoptosis, where the anti-apoptotic genes Bcl2 and Bcl2l12 were found to be direct targets. Again, a downregulation of the miRNA (and thus de-repression of the anti-apoptotic genes) is consistent with what we know about *T. gondii* interference with host pro-apoptotic pathways. The timing of some miR-125 family members is curious with some appearing to be biphasic.

The role of miR-155 in cellular processes is simultaneously well-studied but again the many layers of regulatory complexity mean that the results are not always very clear. This miRNA has been shown to be upregulated by inflammation (especially IL6) (169) but its downstream effects appear to have more to do with cancer metabolism, through an intriguing cascade effect. Transfection with the miRNA was found to result in an increase in Hexokinase 2 protein levels, but, rather than being a direct target of miR-155, it emerged that miR-155 repressed miR-143 and this latter miRNA was found to be a repressor of HK2 (170). Indeed, miR-155 was one of the miRNAs looked at by Cannella et al in the brains of mice. As I did (Figure 5.35), they found it to be upregulated by both strains (171).

The other miRNA identified in that study was miR-146a. In my dataset, though it was statistically-significantly upregulated by all strains, it is clear that the miRNA's upregulation at high MOI ME49 is far greater than the other strains and conditions (Figure 5.36).

Another interesting miRNA to consider in my common core dataset is miR-199, which was downregulated by both strains, at both times (Figure 5.37). This is a hypoxia-related miRNA, and in fact *Hif1a* appears to be one a direct targets (172). The presence of HIF1A has been described as being required for normal growth of *T. gondii* during hypoxia[11] but the mechanism

[11] Though the data suggest this is also true at normoxia, see Chapter 6

for the stabilisation of HIF1A is much less clear. Signalling through the ALK4,5,7 system was shown to be a possible route (173) but the exact mechanism remains unknown. In Spear et al's hands, transcript levels of *Hif1a* did not rise upon infection at normoxia[12] which is consistent with targeting by a miRNA. Thus it is possible that at least part of the stabilisation of HIF1A is via the downregulation of miR-199a.

Finally, I opted to look at miR-23a and miR-23b, given that they are known (suppressed) targets of MYC (164) and that they are known to suppress the expression of glutaminase (see **Chapter 6**). Thus, the expectation was that the overexpression of MYC known to occur during infection would suppress their expression here. Confoundingly, the opposite was true for miR-23b, while miR-23a had an unclear profile, seemingly not very different from the uninfected sample (Figure 5.39). It will be interesting to see what other mechanisms might control the expression of these miRNAs during infection and how that impinges on the expression of downstream genes like glutaminase.

[12] Though it did in mine, see Chapter 6

VI. Effects of *Toxoplasma gondii* Infection on Host Cell Metabolism

6.1 Introduction

6.1.1 *Toxoplasma gondii* infection and host cell gene expression

Large-scale efforts have been conducted in characterising the genome and transcriptome of *Toxoplasma gondii* itself, using 5-prime SAGE tags, microarrays and, of course, RNASeq (including ChIPSeq). Many of these studies have looked at either the difference between strains or at how expression of parasite genes changes as differentiation proceeds from tachyzoite to bradyzoite. These studies have largely been collected and published on ToxoDB which, at the time of writing, hosts some 137 datasets characterising the parasite and closely related species (such as *Neospora caninum*) and of these, 49 involve transcript or protein expression of *T. gondii* (174). In contrast, only three such datasets exist for HostDB[13], the analogous database for host cell infected with the parasite, though admittedly this database in still very much in beta (175). While one of these studies describes host cell transcriptomic changes in response to infection by 25 different strains, the analysis of these data has focussed entirely on parasite gene expression (12). A large-scale proteomics study is also cited but again, the analysis is entirely parasite-centric (176). Of course that is not a true reflection of the study of host-cell transcriptomics in response to infection; a great many have looked at host cell gene expression changes in response to infection by *T. gondii*.

An early time course study was conducted by Blader et al (177) who found a roughly biphasal transcriptional programme elicited by *T. gondii* infection. This experiment looked at HFF cells infected with the Type II PDS

[13] The current URL for HostDB is: http://beta.hostdb.org/hostdb.b28/showApplication.do

strain (an ME49-clone), and used arrays to assay host cell gene expression at 1, 2, 4, 6, and 24h following infection or mock-infection. The authors found that, early in infection (1-6h), pro-inflammatory immune-related genes were differentially-expressed, including chemokines, cytokines (such as *Il6* or *Il1b*). At 24h following infection, genes involved in glycolysis were highly upregulated, but those involved in the citric acid cycle (TCA) were not. Cholesterol biosynthesis was also upregulated, especially via mevalonate metabolism while, as the authors note, *T. gondii*'s method of acquiring sterols from its host is thought to be by subverting LDL-trafficking in the host. It may be then that parasite scavenging depletes host stores which must then be replenished via the mevalonate pathway. Unfortunately, probably due to the relatively high MOI employed in this study (up to 10), the researchers did not follow infection beyond 24h.

It is clear, however, that timing has a large impact on host cell gene expression. A 2006 study by Okomo-Adhiambo et al (178) profiled host cell gene expression via microarray in porcine kidney epithelial cells (PK-13) over a time course of 72h. To achieve this, the researchers used an unprecedentedly (to me) low MOI (1/6), and employed a temperature-sensitive derivative of RH – TS-4 – which grows more slowly at 37° C. PK-13 is a relatively rarely used cell line[14] and currently has a note on the supplier's site (ATCC), which states that this particular cell line "is neither produced nor fully characterized by ATCC." (179) Their choice to use GAPDH as a control housekeeping gene for normalisation is also somewhat unfortunate, given known transcriptomic and proteomic evidence of its upregulation during infection (177, 180), especially when the given justification for this choice was based on 'normal' kidney epithelial cell expression rather than *T. gondii* infection[15].

[14] While obviously not comprehensive or definitive as a test, a search for the terms "PK13" or "PK-13" on 16/08/2016 yield only three PubMed references.

[15] Despite these data, *Gapdh* is a commonly-used reference in *T. gondii*-host interaction studies.

Nevertheless, a number of pathways were identified in this study as being differentially-expressed, including early upregulation of pro-apoptotic factors and late-stage downregulation of TCA-related genes. The authors attempt a comparison with the earlier Spear study (177) but do not make a case for either overwhelming similarity or difference that may have arisen from either the difference in parasite strain (Type II versus Type I), host cell species (human versus porcine), MOI (5-10 versus 1/6) or timing. It is difficult therefore to make generalisations about the timing of host cell responses to infection using these data.

The early transcriptional (largely immune-related) response to *T. gondii* infection was further investigated in 2007, when Kim et al used microarrays to look at the effect of IFNG on infected cells (181). The type II Prugniaud strain was used to infect HFFs and the cells were treated with IFNG 18h post infection (or mock-infection). Remarkably, unlike the uninfected controls, where most genes known to be transcriptionally IFNG-responsive were differentially-expressed following treatment, the infected cells did not show IFNG-based dysregulation. Thus, the use of large-scale transcriptional profiling paved the way for a greater understanding for how the parasite is able to survive a robust host cell immune response and persist through to bradyzoite differentiation. Microarrays were used more recently in a similar manner – focussing on a specific, rather narrow pathway or host cell phenomenon – to identify the MYC-related genes that appeared upregulated upon infection of RAW macrophages with RH for 24h (21).

Microarrays have also been used to explore the difference in virulence between the three canonical strains, by looking at variations in gene expression that differed between types II and III, and then examining host cell effects that segregated in the progeny of the type IIxIII cross. This work was crucial in determining the role of ROP16 and STAT3/STAT6 activation in the virulence differences between the types (182). Broadening the range to 29

different strains (including non-canonical ones), Melo et al performed RNASeq on bone marrow-derived macrophages infected at a variety of MOIs, for 20h. Again, this study was aimed at clarifying the differences between the strains – such as the activation of a Type I interferon response by 'atypical strains' – rather than on the commonalities of *T. gondii* infection (183).

Another common use of large-scale transcriptomics in *T. gondii* infection is when studies concentrate on the difference in host cell response between infection with mutant or wild type parasites, as was done in the case of GRA15 to identify a potent effect on host-cell NFkB and NFkB-regulated genes (17). A similar study – albeit RNASeq, rather than array-based – is currently underway by Bo Shiun Lai (personal communication), seeking to characterise the role of ROP13 on host cell gene expression.

In vivo transcriptomics have tended to use mice as the animal model of choice and these studies have focussed on particular organs. For instance, in 2016 He et al performed transcriptional profiling of mouse livers following a six-day infection by the PYS strain (184). Perhaps unsurprisingly, several immune response pathways were upregulated, such as IFNG and IL1B. Given the liver's function in drug metabolism, the authors of the study concentrated on the potential implications of infection on adverse drug reactions. A proteomic analysis of infected mouse livers followed, from the same laboratory (185). Interestingly, they reported a downregulation in fatty acid metabolism, via an apparent suppression of the peroxisome proliferator-activated receptor (PPAR) signalling pathway.

The same laboratory also conducted transcriptional profiling studies in spleens from mice infected with the Type I strain RH (186). This revealed a number of changes to splenic organelles, including a generalised downregulation of mitochondrial genes.

In terms of metabolism, curiously, while hexokinase 2 (*hk2*) was found to be upregulated (suggesting to me an increase in glycolytic processes), so was

pyruvate dehydrogenase kinase, which normally acts to inhibit the entry of pyruvate into the citric acid cycle (TCA) by inhibiting pyruvate dehydrogenase. The authors did not further examine the metabolic impact of RH infection in the spleen, choosing instead to focus on endolysosomal processes and antigen presentation.

An obvious limitation of microarrays to study gene expression is that they can only profile what has been spotted on them. RNASeq thus provides a huge advantage in that it assays what has been transcribed and so interpretation of the results then becomes a question of computational power and the quality of existing annotation. A large-scale RNASeq study has been conducted in porcine PK-15 (kidney epithelial) cells: a time course of infection using RH, at an MOI of 5 . In my hands, that level of MOI with a Type I strain almost always causes host cell lysis shortly after 24h, and so it is unsurprising that the researchers concluded their time course at 24hpi. The fact that the researchers simultaneously profiled parasite and host transcriptomics is rather unique, although no attempt was made to correlate the two organisms' expression profiles. The researchers found that the largest difference in parasite gene expression occurred in the first few hours of infection, and that large-scale host cell transcriptional changes were comparatively delayed. The main thrust of the analysis here, however, was based on GO-term analysis (which yielded such revealing terms as "cellular processes" being enriched). The KEGG-based analyses are far more informative, with specific pathways being described as enriched. Among those were several cancer-related pathways, along with "p53 signaling pathway" which raise interesting questions about the modulation of the host response that may be related to tumorigenic processes, especially given the known interaction of parasite GRA16 with this pathway (19). However, these issues were left unexplored in that study.

In this chapter, I seek to investigate host cell responses elicited by two strains of *T. gondii* – RH and ME49 – in a global manner. Given the plurality of conditions (host cell type, strain, MOI, timing of infection) used in previous host cell transcriptomic studies, I concentrate here on the cellular pathways and processes that appear to be shared, regardless of strain, MOI or, to an extent, timing of infection. I then focus on a particular feature that emerges as highly significant – that of aerobic glycolysis – and show additional functional results that support the hypothesis that infection with *T. gondii* elicits a remodelling of host cell metabolism towards a cancer-like, Warburg phenotype.

6.2 Methods

6.2.1 RNASeq

Thirty-three RNA extractions were performed, as described in Chapter 5, dried as in **5.2.2-5.2.3** and shipped to the sequencing facility at KAUST, where libraries were prepared using Illumina's TruSeq stranded mRNA (187) protocol, and where they were sequenced on a HiSeq 2000. Libraries and sequencing were again performed by Mr Abhinay Ramaprasad.

I aligned the 33 RNA libraries to the *Mus musculus* (Ensembl GRCm38) using TopHat2 (188), which uses the alignment algorithm of Bowtie2 (119) and then puts resultant alignments through splice-site analysis, using the Ensembl annotation GTF file corresponding to the GRCm38 annotation. I employed a mate-inner-distance of 80bp, and the default value for mate standard deviation (as recommended by Abhinay Ramaprasad, personal communication).

The percentage of each library that aligned in concordant pairs is shown in Figure 6.1.

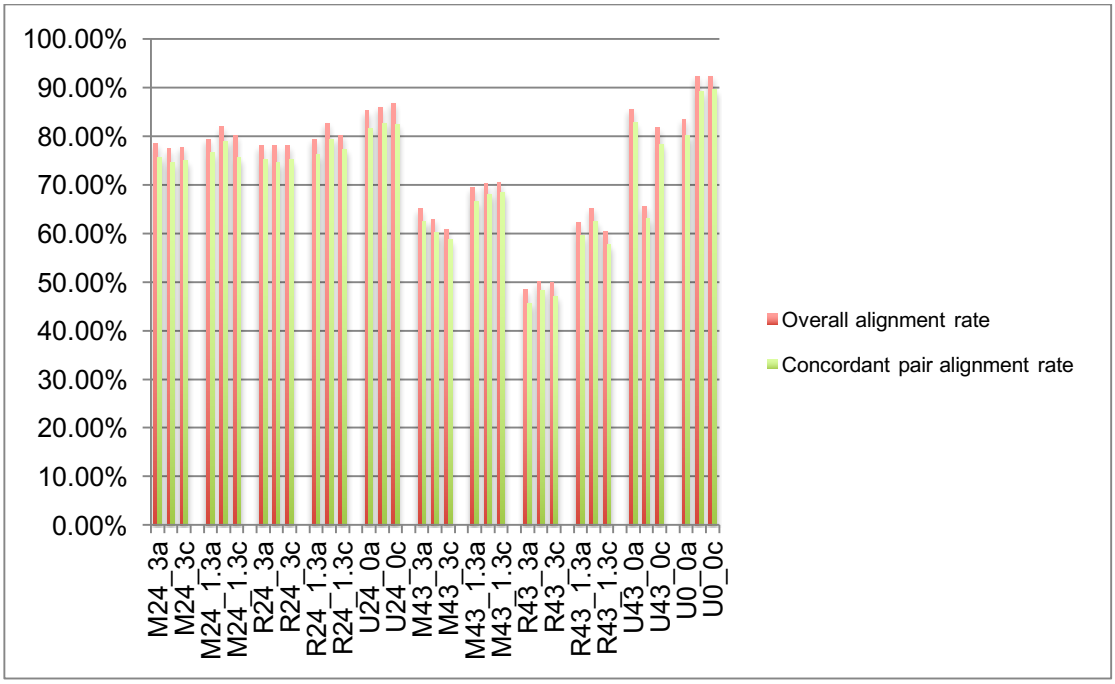


Figure 61: Concordant and overall alignment rate of all 33 samples. NB. Samples are labelled StrainInitialTime_MOI_replicate, e.g. R43_3b is the second biological replicate of cells infected with RH, MOI 3 for 43h while U24_0a is the first biological replicate of uninfected cells at 24h.

Tables of counts were generated using the htseq-count function of the Python library HTSeq (189), which tallies up the reads mapping to a particular feature (genes, as provided by the Ensembl GRCm38 GTF annotation file). I performed the next steps, the differential expression analysis, using EdgeR (140). As suggested by the EdgeR User Manual, I filtered the libraries for genes that a) did not exhibit zero expression across all samples, and b) that corresponded to greater than 1 count-per-million in at least 3 samples. I then performed TMM normalisation, as described in **Chapter 5**.

Principal Component Analysis of my 33 samples following normalisation shows good separation of the different conditions along the second dimension (Figure 6.2).

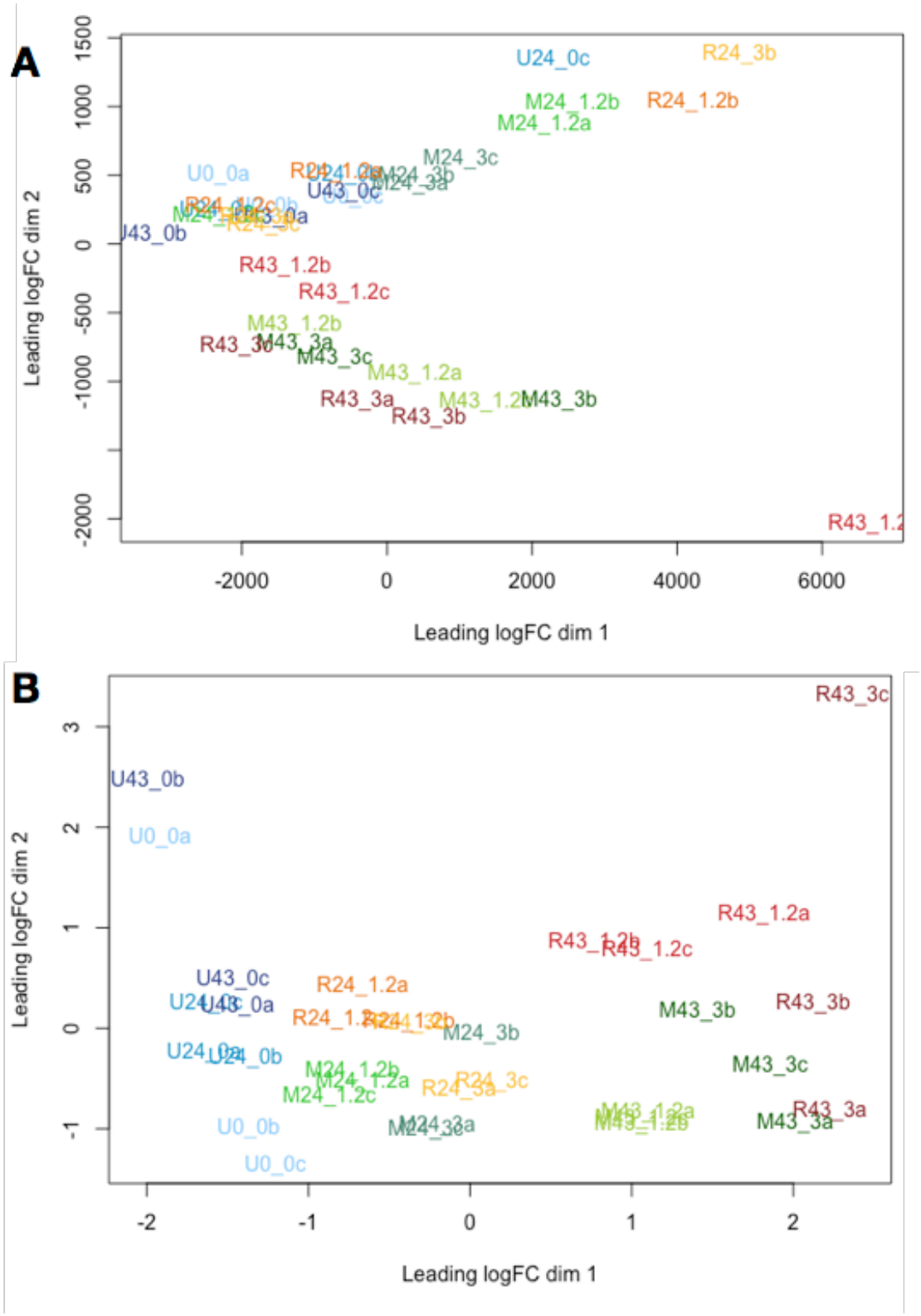


Figure 6.2. Principal component analysis of A) unnormalised and B) TMM-normalised libraries. Sample labels as previous.

6.2.2 KEGG Pathway Enrichment

After identifying genes that were dysregulated either in comparison to the control or to time, I then performed an enrichment analysis of the KEGG pathways contributed to by the genes in each of these profile groups, using the STRING database, version 10.5 (156, 157). This database contains interaction data from over 9.6 million proteins and employs an algorithm to enable the analysis of large gene-lists, akin to the erstwhile popular DAVID (190) tool for microarrays (though it has been updated much more recently than DAVID). Importantly, it allows user-submitted background population lists, which can help avoid bias often found in such enrichment tools (191). For RNASeq it appears that separating gene lists based on directionality can yield more biologically-relevant results (167) and so I assessed functional enrichment in each group separately. I applied a Benjamini-Hochberg multiple test adjustment of 0.05.

6.2.3 Transcriptional Profiles of Individual Genes

Following KEGG pathway enrichment analysis, I selected a number of functionally-related or highly-dysregulated genes and plotted profiles of their transcription over time, using MaSigPro (160).

6.2.4 Western Blotting

Given the apparent importance of *Hif1a* to several of the dysregulated genes, the next step was to perform western blots on key metabolic or cancer-related genes, using cell lysates from HIF1A-KO or HIF1A-WT MEFs infected or mock-infected with ME49. The conditions of infection were MOI 3, for 43h.

6.2.5 Lactate Assay

I performed lactate assays on extracellular medium taken from NIH/3T3 cells infected (or mock-infected) with ME49, at an MOI of 3 over a time course.

6.3 Results

6.3.1 Pairwise Differential Expression

I first examined the RNASeq in a parwise fashion, comparing the infected strains at each time point to the corresponding uninfected time point. I used EdgeR (140) to calculate which genes were differentially-expressed in each of these conditions, at a Bonferroni-Hochberg adjusted p-value of 0.01. For the RH strain samples, an MOI of 1.2 was used, which, as shown in Chapter 5, corresponds (in terms of parasite infection levels) to the ME49-infected samples at an MOI of 3. Several genes were found to be differentially-expressed between the infected and control samples; plots of the log-fold changes are shown in Figure 6.3.

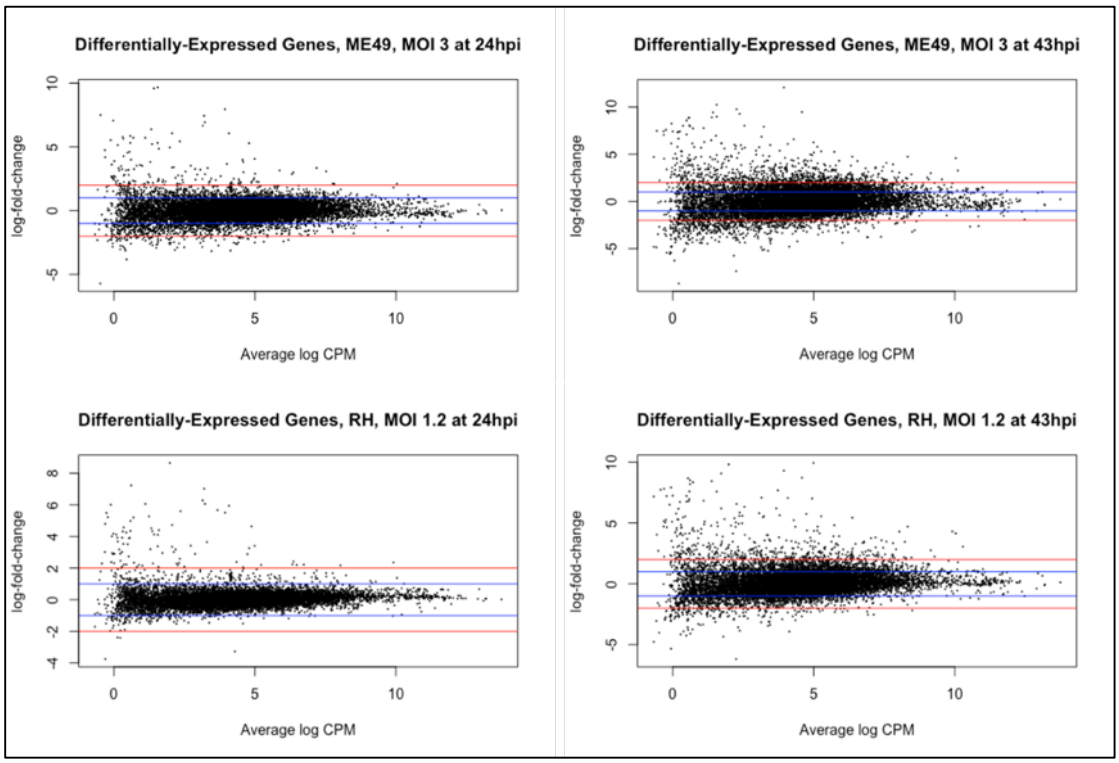


Figure 6.3. Expression plots for each of the infected samples. For each infected sample, differentially-expressed genes, with respect to the uninfected control at the corresponding timepoint were plotted using R. Red lines represent a $\log_2$fold change of 2, blue lines represent a $\log_2$fold change of 1.

As expected, there appear to more many more differentially-expressed genes at the later time-points of infection; this is explored further in **6.3.2**, which looks at differences across the time course as a whole. The full lists of differentially-expressed genes at each time point between each infection condition and the corresponding control samples are in Supplementary Folder 1. The number of differentially-expressed genes in each pairwise comparison are shown in Table 6.1 and Figure 6.4.

Table 6.1. Differentially-expressed genes in each infected sample, as compared to the uninfected sample at the corresponding time-point. DE genes were filtered for an adjusted p-value of less than 0.01 and a $\log_2$ fold change of at least 1.5

Strain and MOI	Upregulated at 24hpi	Upregulated at 43hpi	Downregulated at 24hpi	Downregulated at 43hpi
ME49, MOI 3	132	516	188	393
RH, MOI 1.2	104	461	4	195

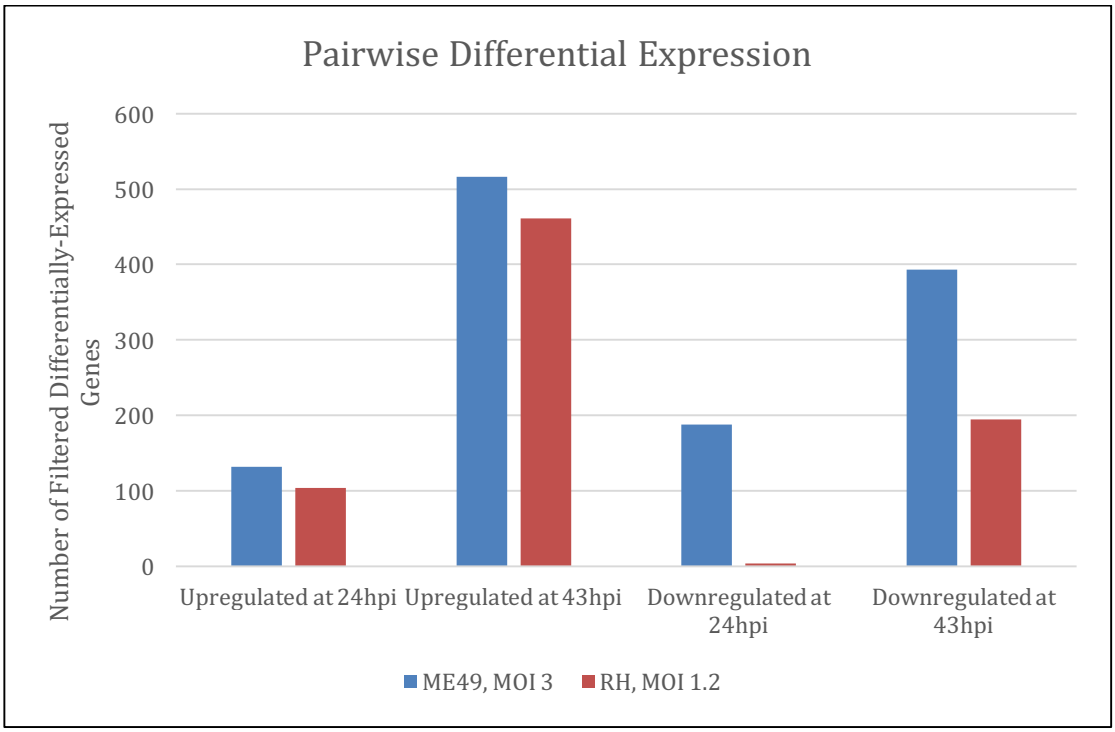


Figure 6.4. Differentially-expressed genes at 24hpi and 43hpi.

Differential expression was calculated between the infected samples and the corresponding, time-matched uninfected controls. They were filtered on an adjusted p-value of 0.01 and a $\log_2$FC of 1.5.

In order to then better understand the interactions of these differentially-expressed genes, I constructed venn diagrams to uncover patterns between the different strains and time-points (Figures 6.5 and 6.6). Given that more genes were dysregulated at the later time point, regardless of the strain used, it could be feasible that those dysregulated at 43hpi include those dysregulated at the earlier time-point. However, it does not appear that this is the case: several genes were uniquely dysregulated at 24hpi, which suggests a potentially-'phased' mode of gene expression following infection. The dysregulation of genes across time following infection is further explored further in this chapter.

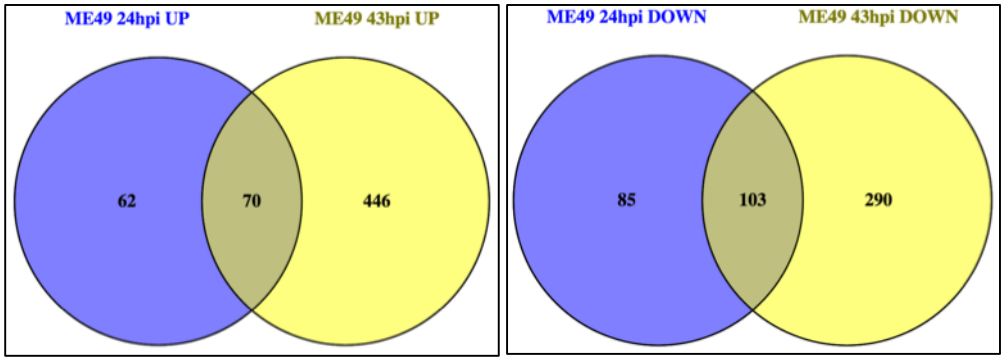


Figure 6.5. Venn Diagram of up- and down-regulated genes following infection with ME49. Differentially-expressed genes from the ME49-infected samples were intersected, based on whether they were up- or down-regulated compared to the uninfected control sample.

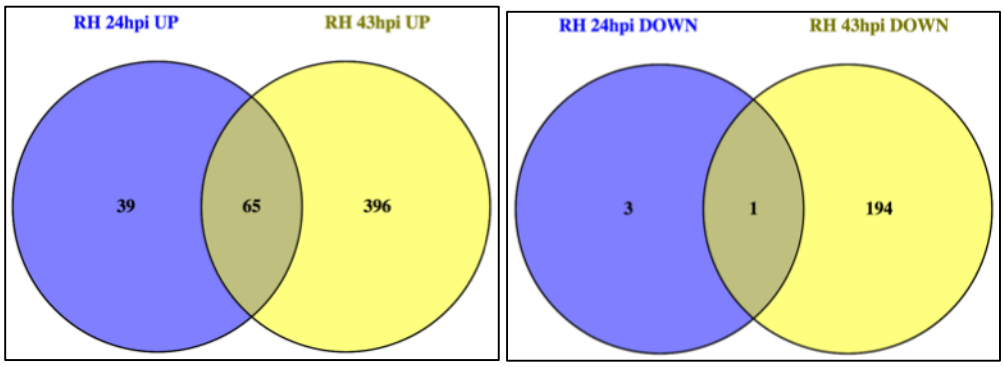


Figure 6.6. Venn Diagram of up- and down-regulated genes following infection with RH. Differentially-expressed genes from the RH-infected samples were intersected, based on whether they were up- or down-regulated compared to the uninfected control sample.

In the RH-infected samples, only a single gene was downregulated both at 24hpi and 43hpi, *lymphocyte antigen 6 complex, locus C2*, a gene that has not been explored in the context of *Toxoplasma gondii* infection.

Given that my interest is in discovering host cell gene expression effects upon infection as a whole, I think intersected the genes that were dysregulated by both strains, at both time-points. This proved possible only for the upregulated genes, given that the sole gene downregulated at both 24hpi and 43hpi by RH was not downregulated (nor upregulated) at any time by ME49 infection. As such, I obtained a 'core set' of 29 genes that were upregulated at both time-points, by both strains – a marker of common infection-regulated genes. These genes are shown in Table 6.2 where it is already apparent that several genes known to be related to *Toxoplasma* infection are represented, for instance members of the NFkB family, chemokine ligands and interleukins.

Table 6.2 All the different infections conditions and time-points were interescted. A list of 29 'common core' genes were found to be upregulated.

Gene Symbol	**Gene Name**
Mmp9	Matrix metallopeptidase 9
Traf1	TNF receptor-associated factor 1
Nfkbia	Nuclear factor of kappa light polypeptide gene enhancer in B cells inhibitor, alpha
Rnf19b	Ring finger protein 19B
Nfkb2	Nuclear factor of kappa light polypeptide gene enhancer in B cells 2, p49/p100
Ccl5	Chemokine (C-C motif) ligand 5
Enpp2	Ectonucleotide pyrophosphatase/phosphodiesterase 2
Dcn	Decorin
Stra6	Stimulated by retinoic acid gene 6
Acpp	Acid phosphatase, prostate
Egr1	Early growth response 1
Nfkbie	Nuclear factor of kappa light polypeptide gene enh
Lif	Leukemia inhibitory factor
Ubxn2a	UBX domain protein 2A

Egr2	Early growth response 2
Mfsd2a	Major facilitator superfamily domain containing 2A
Slc7a11	Solute carrier family 7 (cationic amino acid transporter, y+ system), member 11
Cd200	CD200 antigen
Homer1	Homer homolog 1 (Drosophila)
Dusp4	Dual specificity phosphatase 4
Tmem45a	Transmembrane protein 45a
Il23a	Interleukin 23, alpha subunit p19
Gfpt2	Glutamine fructose-6-phosphate transaminase 2
Cxcl1	Chemokine (C-X-C motif) ligand 1
Saa3	Serum amyloid A 3
Tnfaip3	Tumor necrosis factor, alpha-induced protein 3
Cyp4f40	Cytochrome P450, family 4, subfamily f, polypeptide 40
Dennd2d	DENN/MADD domain containing 2D
Tslp	Thymic stromal lymphopoietin

Using this list of genes, I then examined the relationships between them, using STRING (192), a database of protein interactions. Using the authors' recommended settings, I constructed a network diagram of the genes up-regulated by both strains. The main protein interactions that were found in the 'core upregulated' set of differentially-expressed genes in both strains and both timepoints is shown in Figure 6.7. The major nodes are well-known genes differentially-expressed in infection, including NFkB family-members and known markers of infection such as CCL5 and Interleukin-23. One interesting cluster that emerged from this analysis was that containing the SLC7A11 protein, which is a cystine/glutamate antiporter known to be upregulated in several cancers (193).

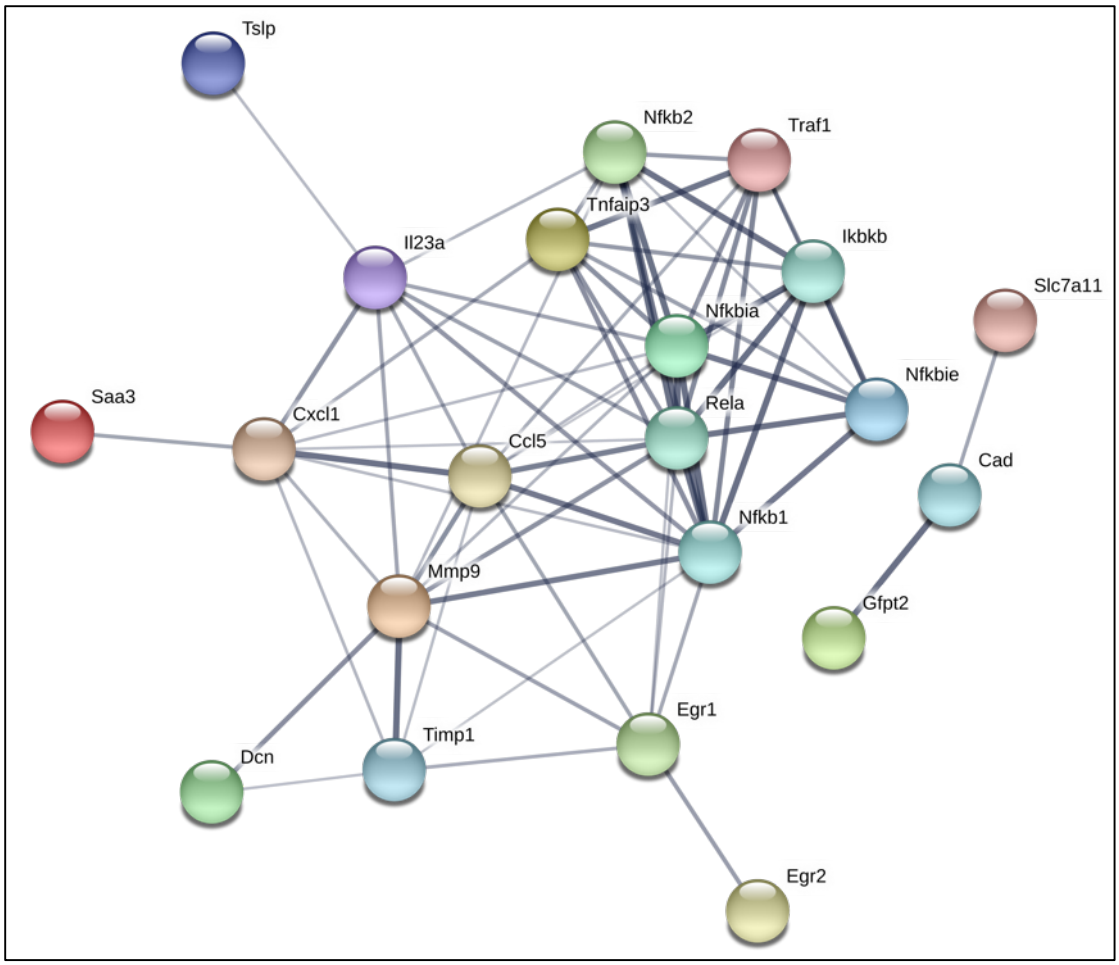


Figure 6.7. Protein interactions of genes found to be differentially-expressed in both strains, at both times. The core set of 29 genes found to be upregulated in all conditions were mapped and their interactions plotted. The colours of the nodes are simply to distinguish them from each other and do not hold any particular meaning.

STRING (192) also performs KEGG pathway enrichment for a inputted set of genes. The enriched pathways are shown in Table 6.3 alongside the false discovery rate for that enrichment. Clearly, there are several expected pathways, such as the NFkB signalling pathway and the TNF pathway. However, some other more curious pathways emerge, such as several that implicate cancer-like mechanisms. These patterns are explored further by addressing the individual gene contributions to these enrichments (**6.3.3**).

Table 6.3. The core set of upregulated genes (upregulated by each strain and at both time points) was analysed for KEGG pathway enrichment using the STRING database. Genes contributing to the enrichment groups are presented in Supplementary Folder 1.

#pathway ID	pathway description	observed gene count	false discovery rate
4668	TNF signaling pathway	7	1.86E-08
4621	NOD-like receptor signaling pathway	4	0.000116
4064	NF-kappa B signaling pathway	4	0.000363
5169	Epstein-Barr virus infection	5	0.000363
4060	Cytokine-cytokine receptor interaction	5	0.000773
5134	Legionellosis	3	0.00249
5203	Viral carcinogenesis	4	0.0045
5166	HTLV-I infection	4	0.0122
4630	Jak-STAT signaling pathway	3	0.0217
5020	Prion diseases	2	0.0217
5161	Hepatitis B	3	0.0217
5200	Pathways in cancer	4	0.0217
4062	Chemokine signaling pathway	3	0.0322
5168	Herpes simplex infection	3	0.04

As well as this core set of commonly dysregulated genes, I also wished to explore some of the differences between the strains. I thus then filtered my gene lists to retain only those that were *unique* to either strain, *either* time point, Figures 6.8.

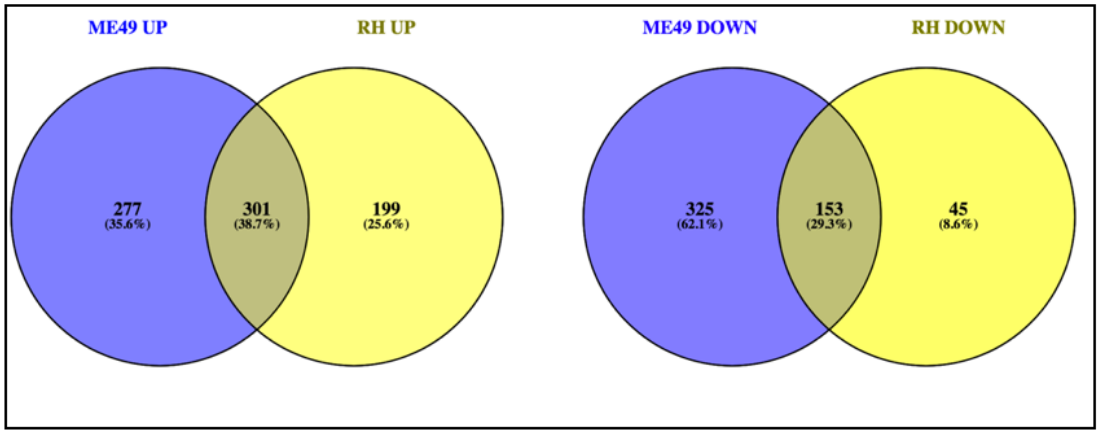


Figure 6.8. Genes that were dysregulated *at either time point* were intersected, to obtain lists of dysregulated genes unique to each strain.

Overall, it appears that more genes were uniquely dysregulated in ME49. In the upregulated geneset, 35.6% were unique to ME49, while 25.6% were unique to RH. The situation for the downregulated was even more extreme, with approximately seven times more genes being uniquely downregulated by ME49.

Enrichments were once again performed on each of these gene sets using STRING. I also performed a network analysis, though with higher stringencies, due to the larger number of genes involved. Networks that emerged as being unique to ME49 upregulation are shown in Figure 6.9.

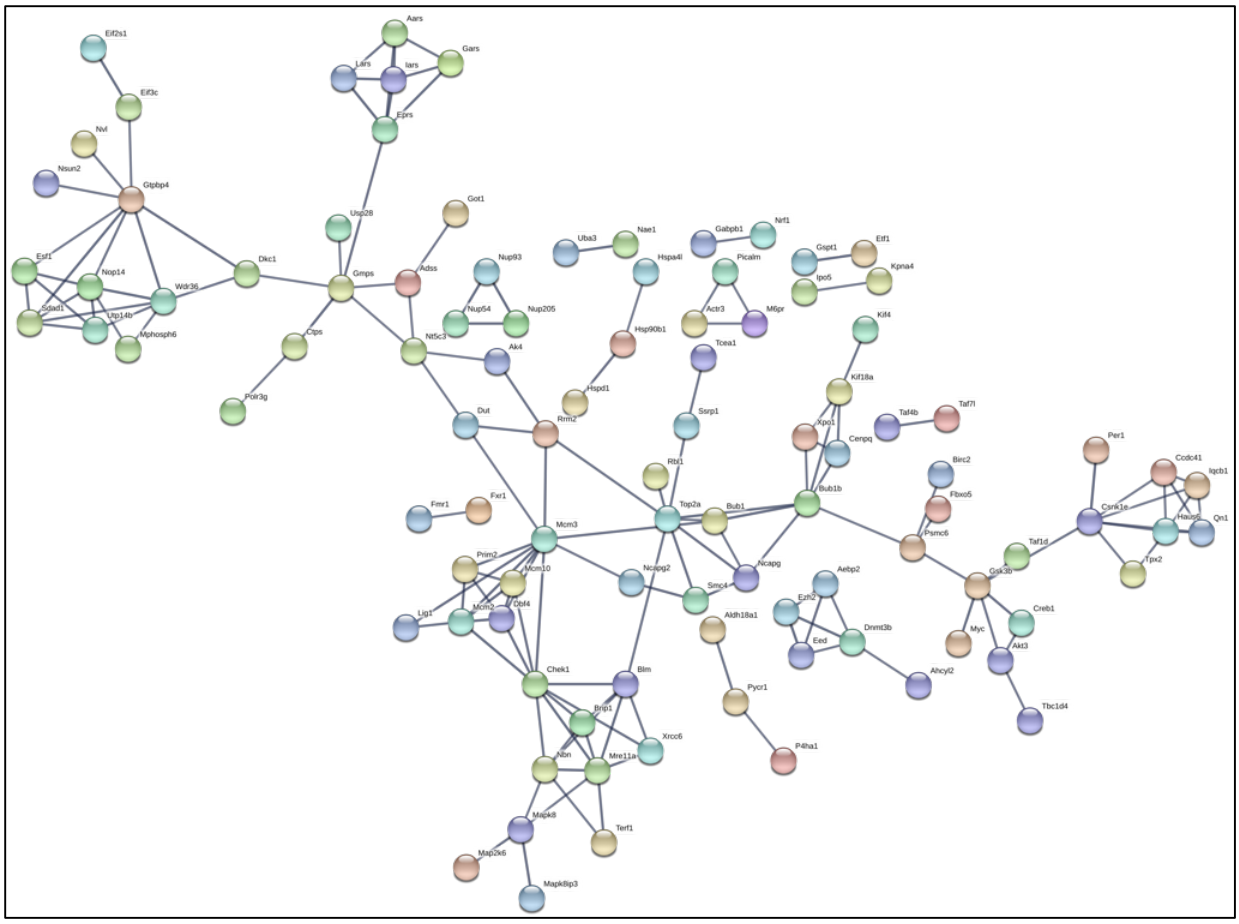


Figure 6.9. Protein interactions of genes found to be differentially-upregulated only in ME49-infected cells, at either time. The 277 genes found to be upregulated only in ME49-infection were mapped and their interactions plotted using STRING. The colours of the nodes are simply to distinguish them from each other and do not hold any particular meaning.

While fewer genes were found to be unique to dysregulation by RH, when I performed a similar network interaction analysis, several more obvious nodes were revealed (Figure 6.10). For instance, a clear JUN-signalling node emerges, with correlations to NF-kB family members. Similarly, the KEGG enrichment analysis also reveals more enriched pathways (Table 6.4).

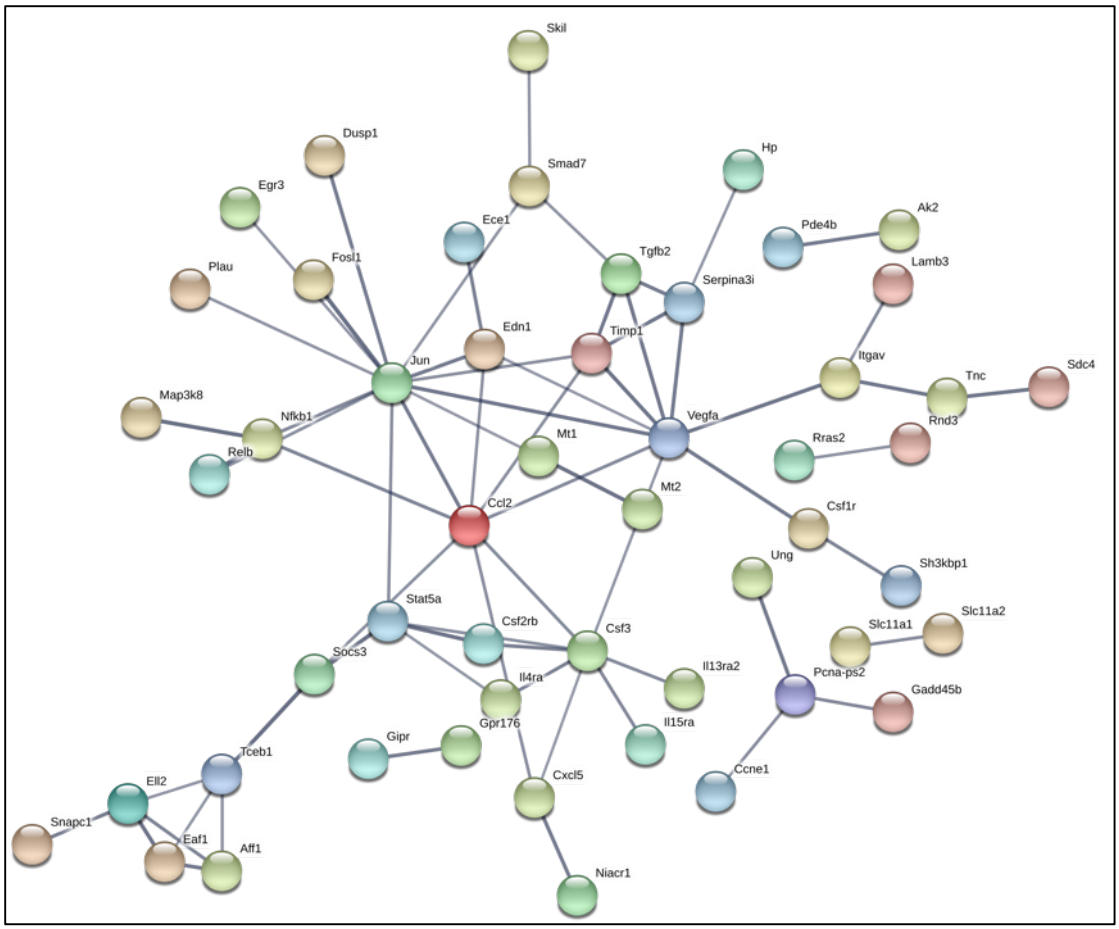


Figure 6.10 Protein interactions of genes found to be differentially-expressed only in RH-infected cells, at either time. The 199 genes found to be upregulated only in RH-infection were mapped and their interactions plotted using STRING. The colours of the nodes are simply to distinguish them from each other and do not hold any particular meaning.

Given the long lists of genes involved in these KEGG pathways, these are shown in Supplementary Folder 1. Table 6.4 shows the KEGG pathways themselves, and whether those pathways are enriched in both strains or in either one of them uniquely.

Table 6.4. KEGG pathways uniquely enriched in infection by either strain

Pathways Unique to ME49	Pathways Unique to RH	Pathways in Common
Ribosome biogenesis in eukaryotes	MAPK signaling pathway	TNF signaling pathway
Colorectal cancer	Cytokine-cytokine receptor interaction	HTLV-I infection

Cell cycle	Rheumatoid arthritis	Pathways in cancer
Epstein-Barr virus infection	Inflammatory bowel disease (IBD)	Osteoclast differentiation
DNA replication	Jak-STAT signaling pathway	MicroRNAs in cancer
ErbB signaling pathway	Mineral absorption	Hepatitis B
RNA transport	PI3K-Akt signaling pathway	
Aminoacyl-tRNA biosynthesis	HIF-1 signaling pathway	
Prolactin signaling pathway	Leishmaniasis	
Endometrial cancer	Renal cell carcinoma	
Thyroid hormone signaling pathway		
Arginine and proline metabolism		
Prostate cancer		
Influenza A		
Purine metabolism		
Hepatitis C		
Pyrimidine metabolism		
Fc epsilon RI signaling pathway		
Biosynthesis of amino acids		
Mismatch repair		
Toxoplasmosis		
Metabolic pathways		
Fanconi anemia pathway		
Progesterone-mediated oocyte maturation		
Small cell lung cancer		
Homologous recombination		
Acute myeloid leukemia		
Proteoglycans in cancer		
Circadian rhythm		
FoxO signaling pathway		
Estrogen signaling pathway		
Toll-like receptor signaling		
Pancreatic cancer		

Interestingly, several pathways were found to be enriched by both strains, despite those enrichments deriving from genes unique to each infection case. Moreover, several recurring pathways, even when unique to a particular strain, refer to metabolic pathways, cancer of various types and, of course, parasitology (Toxoplasmosis, and Leishmaniasis). The component genes of many of these pathways, as well as those highlighted as key nodes in the network analyses are profiled individually in **6.3.3**, where their strain and MOI-specific patterns are further discussed.

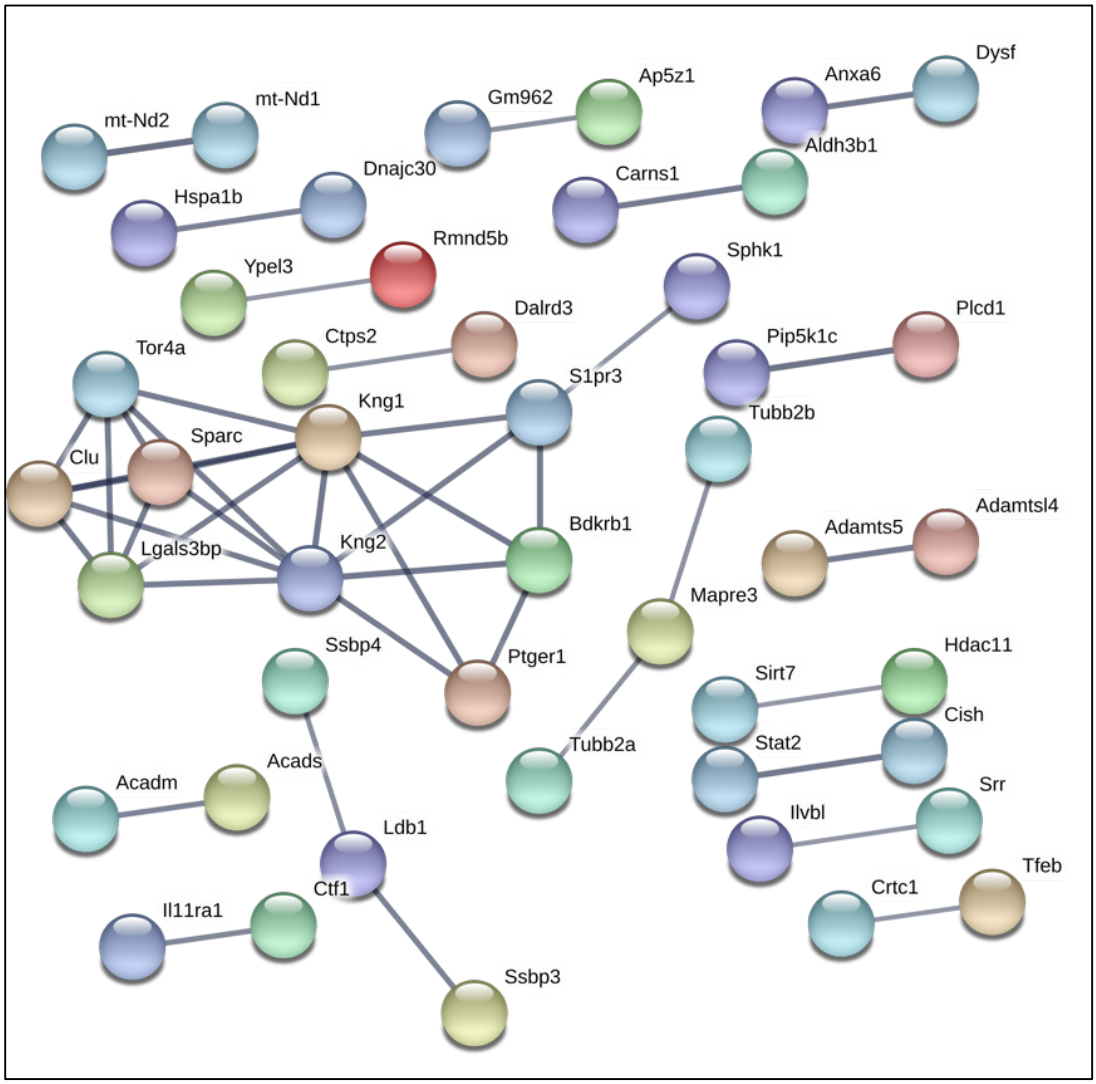


Figure 6.11 Protein interactions of genes found to be differentially-downregulated only in ME49-infected cells, at either time. The 325 genes found to be downregulated only in ME49-infection were mapped and their interactions plotted using STRING. The colours of the nodes are simply to distinguish them from each other and do not hold any particular meaning.

Though a few networks emerged from the STRING analysis of genes uniquely downregulated in ME49-infected cells (Figure 6.11), these did not correspond to functional enrichment via KEGG pathway analysis – no pathways were found to be enriched. That being said, some genes of interest that emerged were *Cish* and *Stat2*, both of which have been implicated in ME49-infection and are discussed further in relation to Figure 6.17. The situation for genes uniquely downregulated in RH was even worse in terms of functional classification. No significant networks of protein interactions emerged, neither did any KEGG pathway enrichments.

6.3.2 Differential Expression over Time

The analysis in **6.3.1** considered each infected sample as compared to its corresponding uninfected control, such that, for example, the sample infected by ME49 for 24 hours was compared to the control. However, given that these infections were performed over time, one can also measure differential expression over the full 43 hours. For this analysis, I again used EdgeR with an adjusted p-value of 0.05. I first looked at genes that were differentially-expressed between infected samples at 24hpi and the zero hour control and then between those in which infection had been allowed to proceed for 43h and the previous (24hpi) time point. Given the extremely large number of genes that emerged as being differentially-expressed over time (more than can be usefully handled by most enrichment algorithms), I opted to select the genes with the 100 lowest adjust p-values at either time comparison to analyse for functional significance, again using STRING for protein interaction and KEGG pathway enrichment.

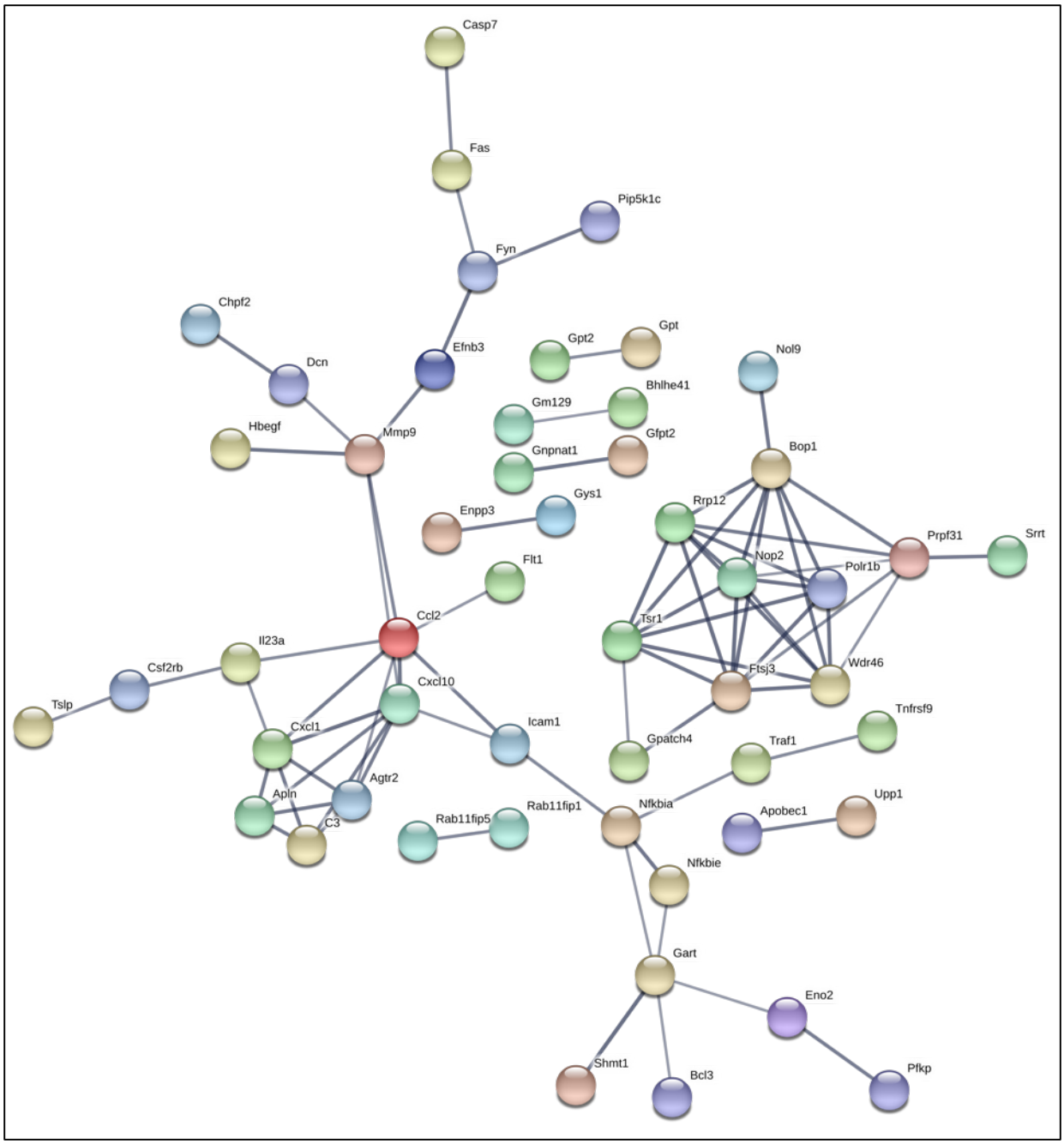


Figure 6.12. A network of protein-interactions from the 100 most significant differentially-expressed genes across time in each ME49-infected samples. The 100 genes from each time comparison (24hpi vs 0h and 43hpi vs 24hpi) were collated and subjected to protein interaction analysis.

Several key nodes that have already been implicated in ME49-infection emerge from the networks shown in Figure 6.12. Notably, members of the NF-kB family are present, as are metabolism-related genes such as *Enolase 2*. Apoptosis-related genes also appear to vary over time in these ME49-infected samples, with a Caspase-Fas association being present in the results. The KEGG pathway enrichment for these same genes provides slightly more context,

where intriguing pathways such as carbon metabolism and TNF signalling are overrepresented. The contributing genes are in Supplementary Folder 1.

Table 6.5 KEGG Pathway enrichment for the top 100 genes from ME49-infected samples *over time*. The top 100 DE-genes across time (at both time comparisons) were examined for pathway enrichment.

#pathway ID	pathway description	observed gene count	false discovery rate
4668	TNF signaling pathway	10	3.92E-06
1200	Carbon metabolism	7	0.00313
4060	Cytokine-cytokine receptor interaction	9	0.0136
1230	Biosynthesis of amino acids	5	0.0208
1120	Microbial metabolism in diverse environments	7	0.022
5134	Legionellosis	4	0.0497

Though the number of genes used for functional analysis of RH-infected was the same as for ME49, more protein-interactions appear in the network (Figure 6.13). As well as the apoptosis-related pathways, a clear cell-cycle cluster appears, containing nodes such as PLK1 and CCNA2. Members of the JAK-STAT pathway also appear, with a connection to the glycolysis-related protein HK1.

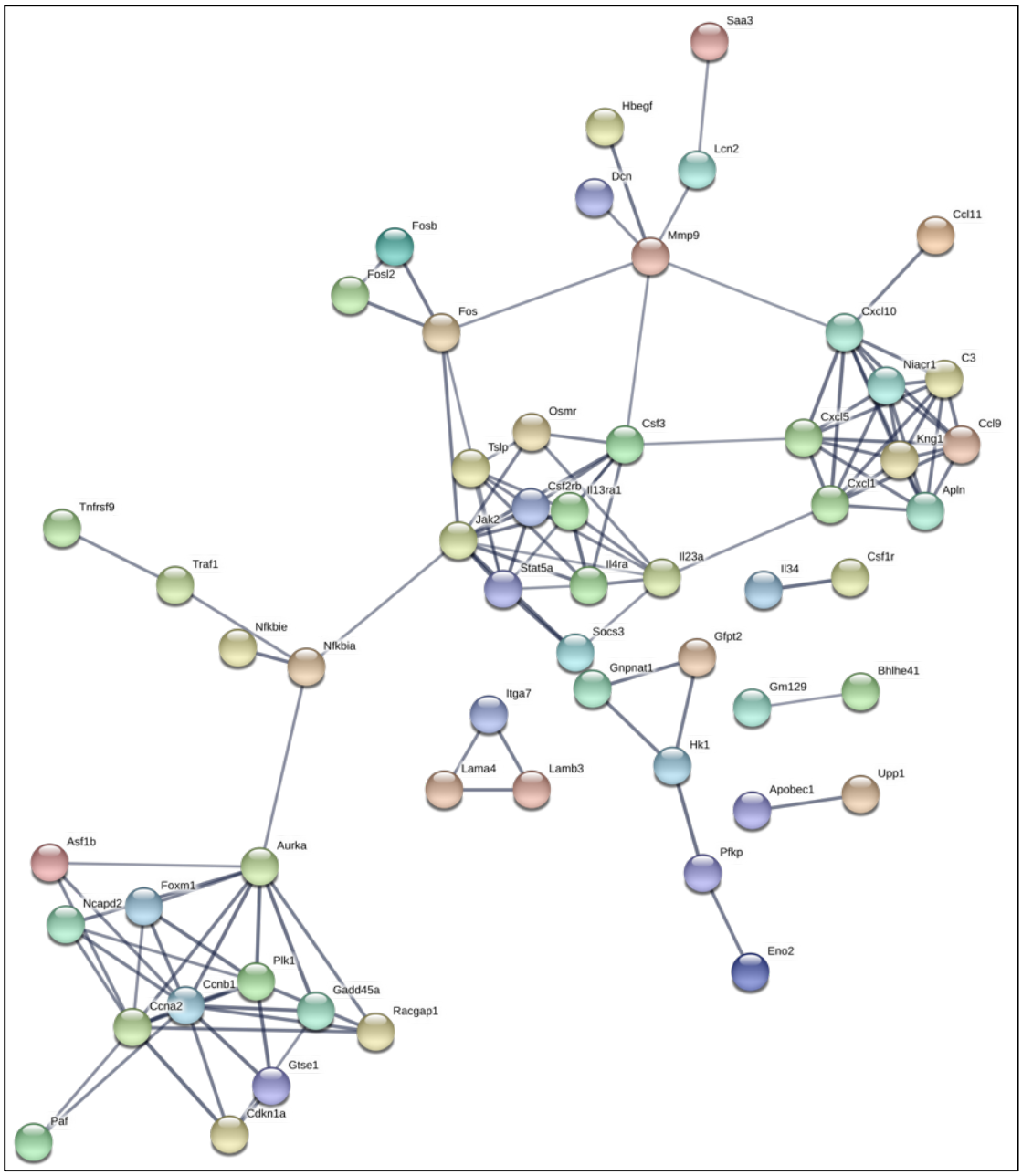


Figure 6.13. A network of protein-interactions from the 100 most significant differentially-expressed genes across time in each RH-infected samples. The 100 genes from each time comparison (24hpi vs 0h and 43hpi vs 24hpi) were collated and subjected to protein interaction analysis.

This is echoed in the KEGG pathways analysis (Table 6.6), where those pathways are clearly enriched for. Additionally, as expected, parasitology-related pathways also emerge, such as that for Leishmaniasis, as well as generalised infection.

Table 6.6 KEGG Pathway enrichment for the top 100 genes from RH-infected samples *over time*. The top 100 DE-genes across time (at both time comparisons) were examined for pathway enrichment.

#pathway ID	pathway description	observed gene count	false discovery rate
4060	Cytokine-cytokine receptor interaction	15	3.20E-07
4668	TNF signaling pathway	9	2.77E-05
4630	Jak-STAT signaling pathway	10	3.13E-05
5200	Pathways in cancer	12	0.000665
4115	p53 signaling pathway	5	0.00713
5161	Hepatitis B	7	0.00713
5168	Herpes simplex infection	8	0.00713
4920	Adipocytokine signaling pathway	5	0.00783
5133	Pertussis	5	0.00783
4380	Osteoclast differentiation	6	0.012
4062	Chemokine signaling pathway	7	0.0137
5203	Viral carcinogenesis	7	0.0235
1200	Carbon metabolism	5	0.0313
4151	PI3K-Akt signaling pathway	9	0.0313
5140	Leishmaniasis	4	0.0313
4110	Cell cycle	5	0.047
4917	Prolactin signaling pathway	4	0.047

6.3.3 Individual Gene Profiles

The pathways signified as being enriched in these analyses appeared to contain several common pathways and genes. Therefore, I profiled them individually, grouping them into functional categories beyond KEGG association (that is, by looking at their interactions in the literature). This section presents the gene profiles with basic context for their inclusion in my examination – I delve more deeply into the way in which they interact in the

context of *Toxoplasma* infection in the Discussion portion of this chapter, **6.4**. In the following profiles, the colour scheme is as shown in Table 6.7.

Table 6.7 Colour scheme for the profiles throughout 6.3.3

Colour	Red	Green	Dark Blue	Light Blue	Pink
Label	Control, Uninfected	ME49, MOI 1.2	ME49, MOI 3	RH, MOI 1.2	RH, MOI 3

Parasitology-Related Genes

The KEGG pathways that were obviously related to parasitology included Leishmaniasis and, of course, Toxoplasmosis. Many of the contributing genes were also contained within other categories (Interferon gamma receptor 2, *Ifngr2*, for instance is discussed in the chemokine/cytokine and receptors section and *Birc2* forms part of the apoptosis pathway). Two interesting cases within this category are the adhesion-related genes Vascular cell adhesion protein 1 (*Vcam1*) and Intercellular adhesion molecule 2 (*Icam1*).

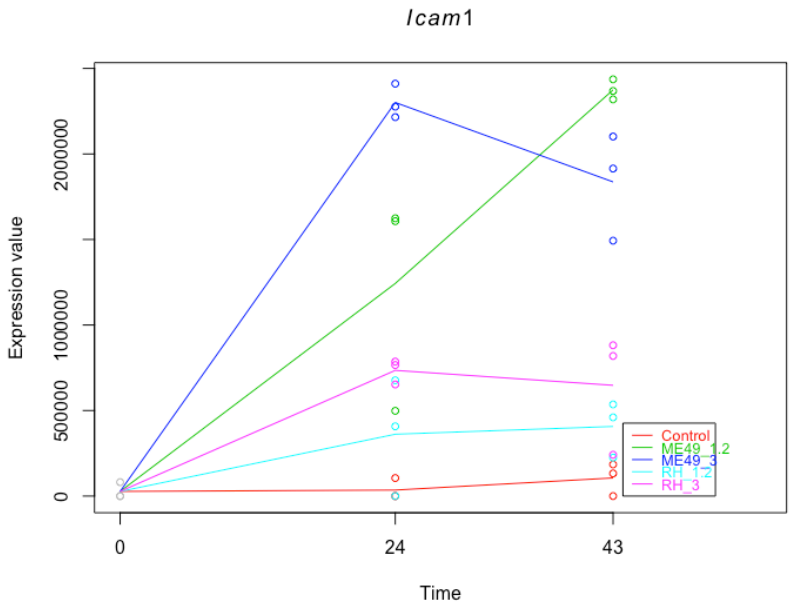


Figure 6.14. Expression profiles of *Icam1* and *Vcam1*.

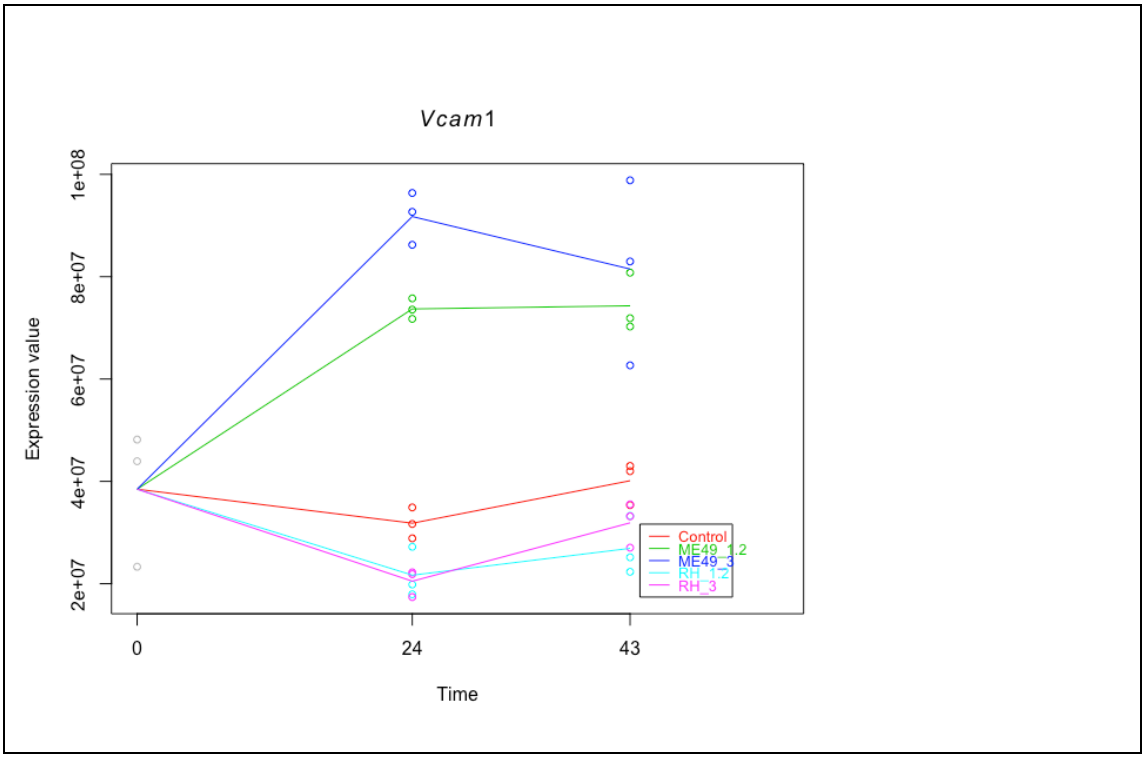


Figure 6.14, continued. Expression profiles of *Icam1* and *Vcam1*.

Host Immune Response

It is difficult to assay the host immune response in non-immune cell culture systems lacking exogenous addition of cytokines – especially in the case of *T. gondii* where, for instance, (IFNG) is such a key player in mediating immunity. *In vivo*, infection by *T. gondii* sets in motion a cascade of proinflammatory processes whereby recognition of the parasite by toll-like receptors (TLRs) or chemokine (C-C motif) receptor 5 (CCR5) lead to the secretion of IL12 (via the NF-kB or MAPK pathways) by dendritic cells. Interleukin-12 in turn recruits NK cells which secrete IFNG and thus begins a programme of immune resistance to the parasite. While of course IFNG itself would not be expected to be detected transcriptionally in my dataset (and indeed it is not, data not shown), its receptor (a heterodimer of IFNGR1 and IFNGR2) is upregulated in all strains, relatively early in infection though *Ifngr1* expression in RH MOI 3 plummets after this time point. Differential expression of the receptor has not been researched in the context of *T. gondii* infection but decreased transcription has been implicated in *Mycobacterium*

tuberculosis-infection as a potential mechanism of immune evasion (194). Conversely, treatment of a human breast cancer cell line (MCF-7) with indole-3-carbinol was found to increase *Ifngr1* and resulted in concomitant augmentation of the IFNG response, showing that transcriptional regulation of this receptor is possible and perhaps underexplored as a means of immune response regulation (195). That I see expression in fibroblasts is consistent with an earlier observation that while IFNG is secreted by NK cells (and T lymphocytes), its receptor is able to be expressed in all cells except erythrocytes (196).

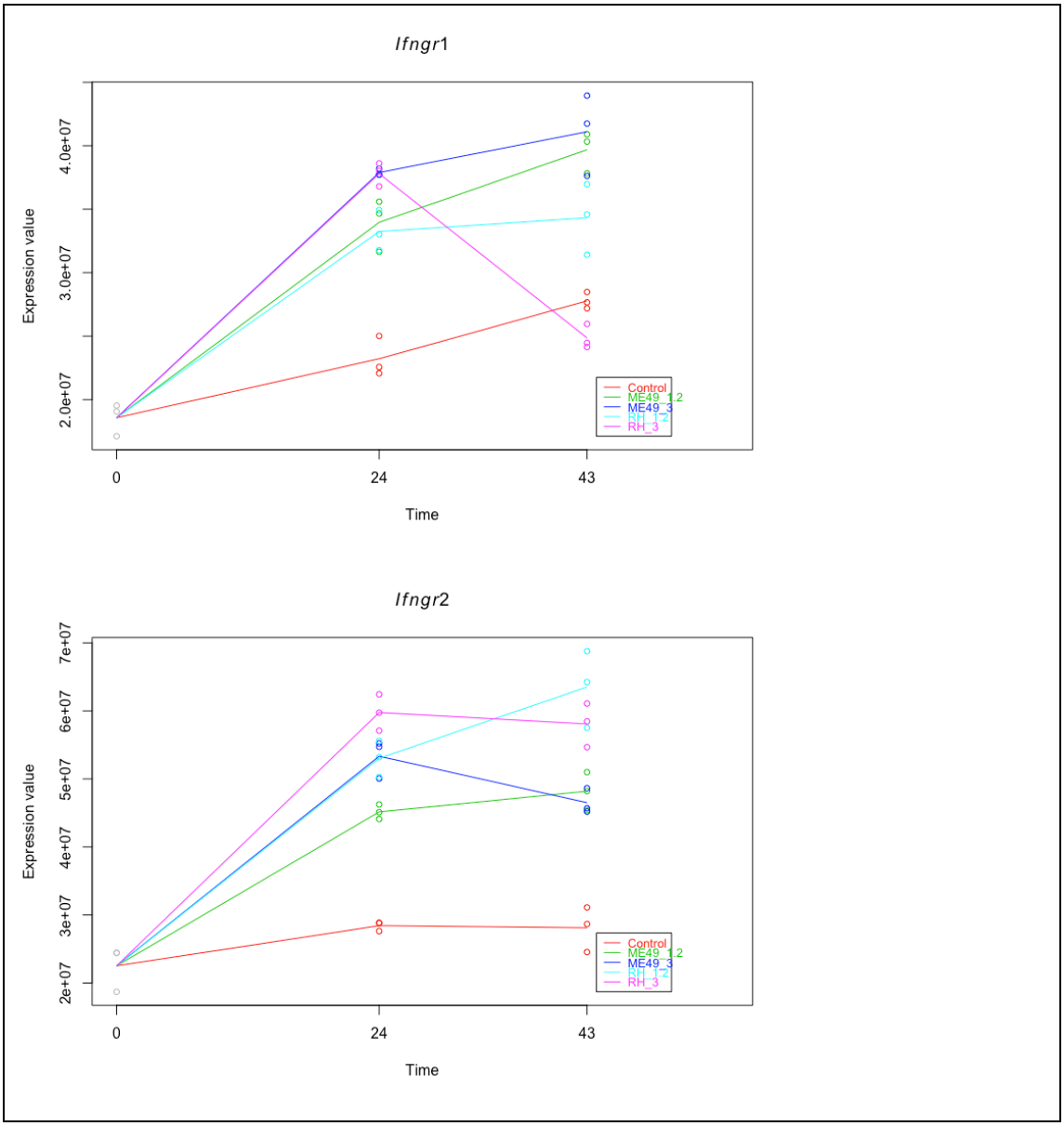


Figure 6.15. Expression profiles of *Interferon gamma Receptor 1* and *2*

Normally, activation of the IFNG pathway then activates a transcriptional programme mediated by STAT1. However, *T. gondii* is able to disrupt this programme, though the mechanism and breadth of how this is achieved appears to be at least in part strain-specific. For instance, it is thought that activation of IRF1 by GRA15 in Type II parasites might induce a specific subset of IFNG-related genes (197). Similarly, the suppressor of cytokine signaling family of proteins – *Socs1*, *Socs3* and *Cish* – have also been implicated in the abrogation of IFNG-responsive gene expression in *T. gondii*-infected cells at least in part in a strain-specific manner (Figure 6.16 and 6.17).

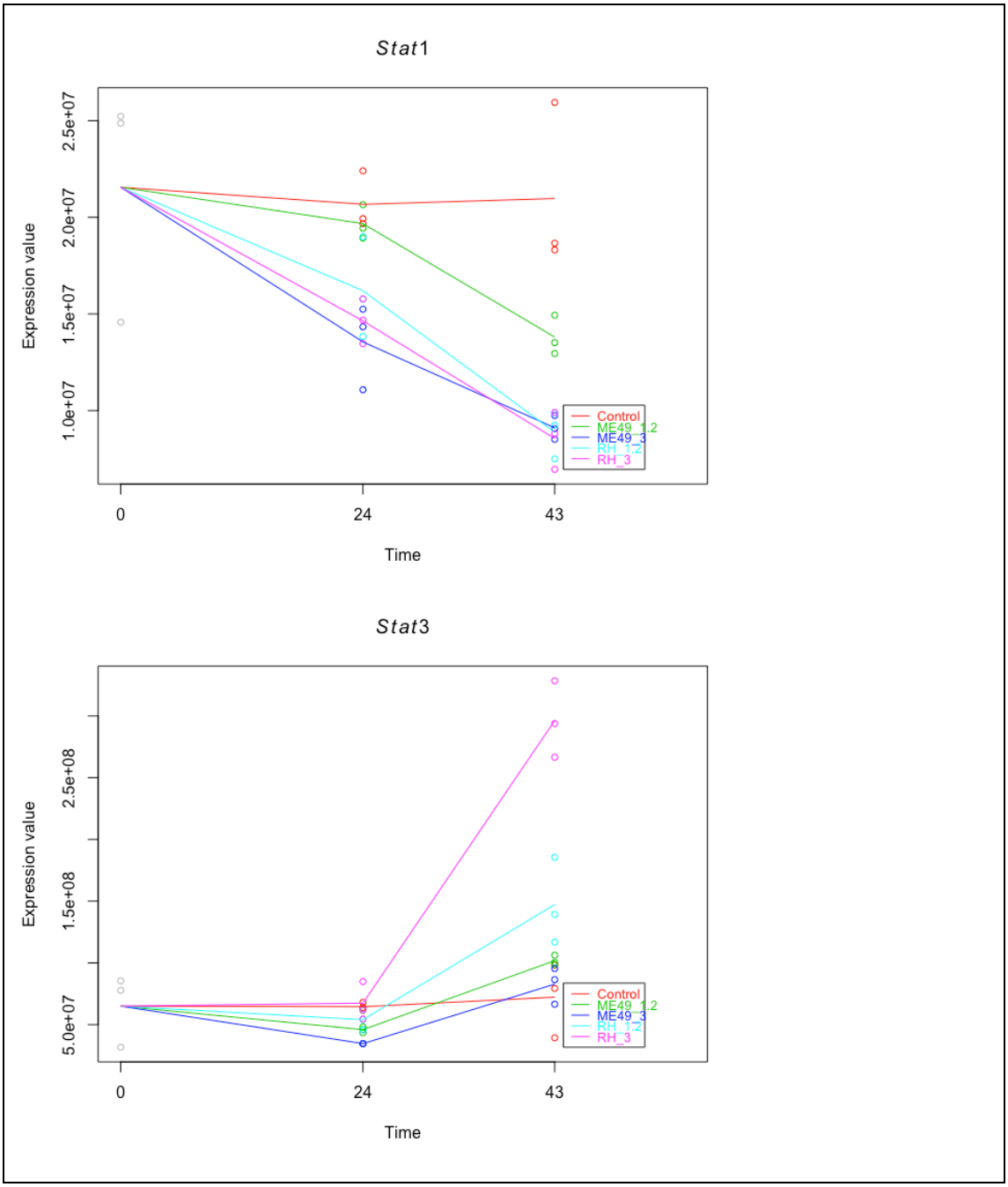


Figure 6.16. Expression Profiles of *signal transducer and activator of transcription 1* and *3*

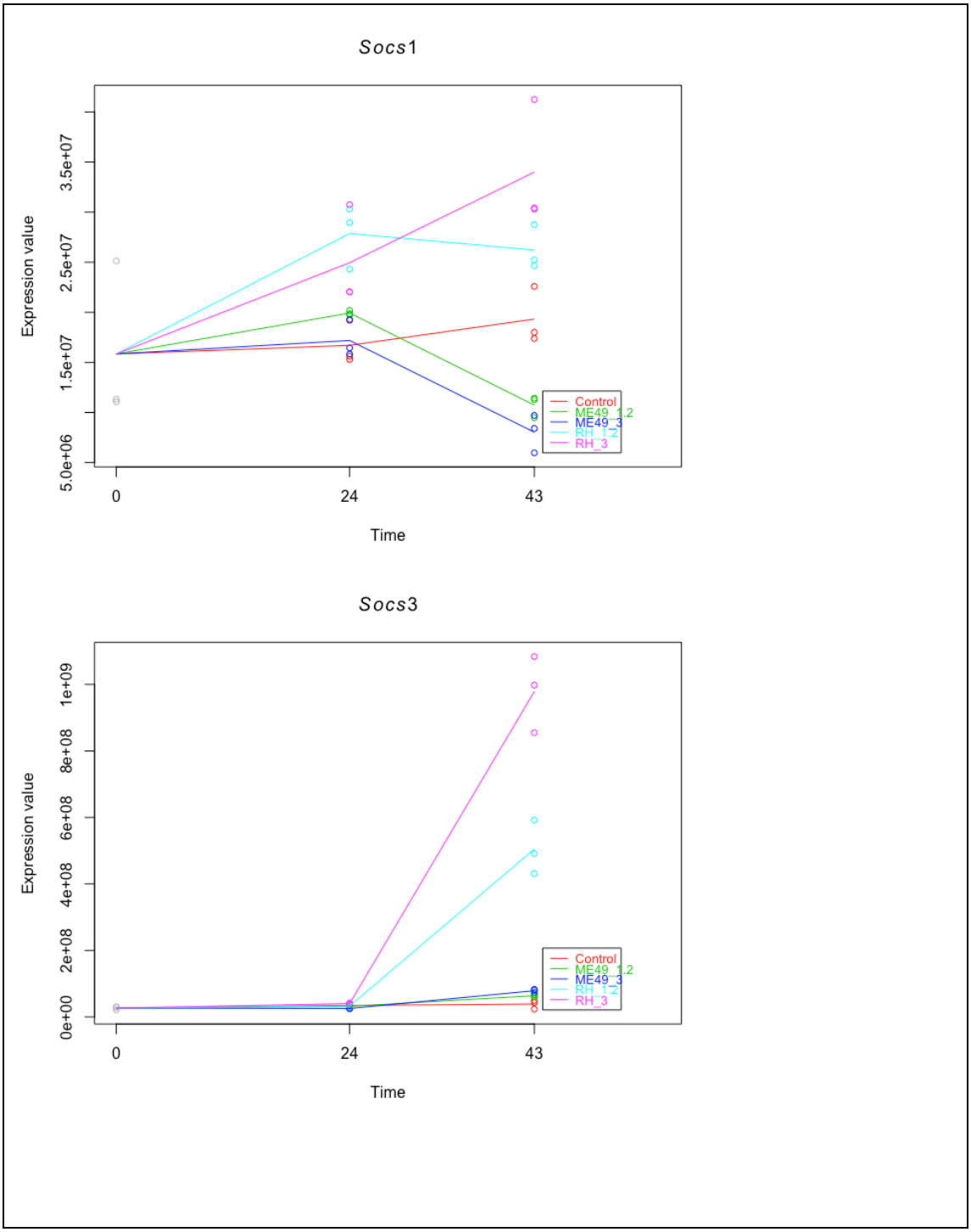


Figure 6.17. Expression Profiles of STAT-related genes

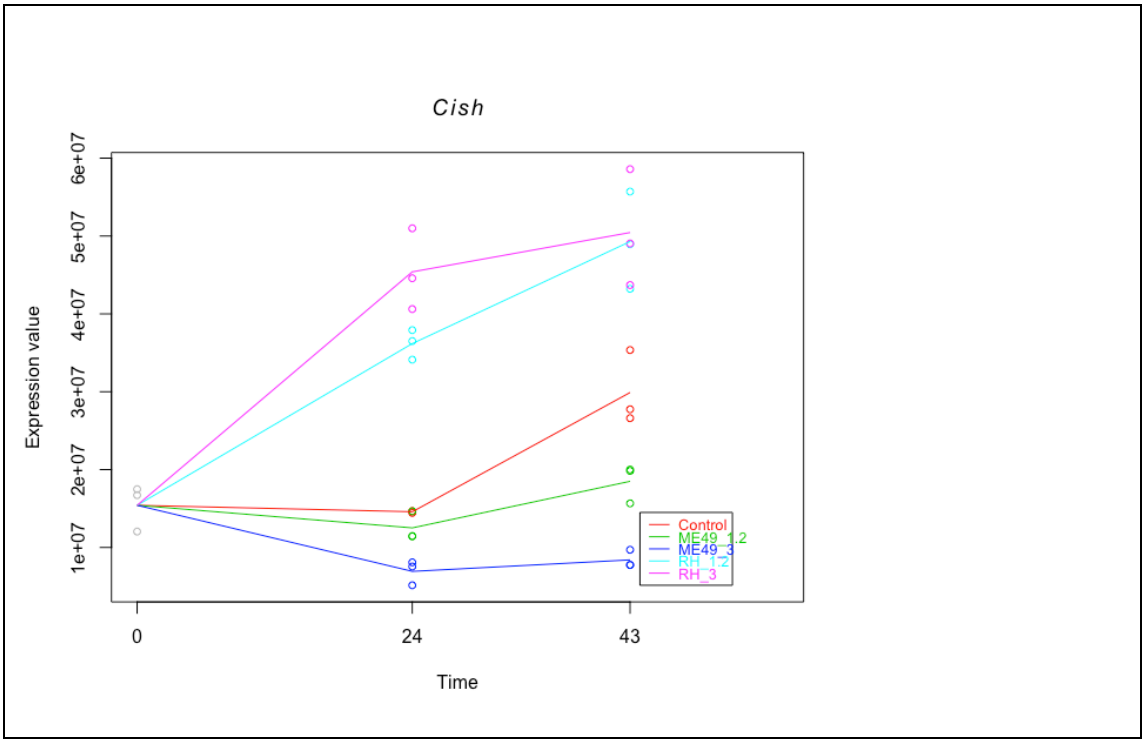


Figure 6.17, continued. Expression Profiles of STAT-related genes

Cell Cycle

When I looked at the list of genes that were dysregulated over time it emerged that several were related to the cell cycle – a process that is also disrupted by *T. gondii* infection. The mechanism of this disruption, and how it fits into host cell metabolism, is discussed in **6.4**. *Cdc25a* appears elevated throughout the time course in all strains, consistent with its role in S-phase promotion; which the parasite also elicits. Perhaps more dramatic is the profile of Cyclin 1 (*Ccne)* exhibiting a strong and sustained upregulation throughout infection. Other cell-cycle related genes that are thought to be modulated by the parasite include *Uhrf1* (148) and *Rb1* which, curiously, functions as a tumour suppressor gene. A gene not usually identified with *T. gondii*-mediated cell cycle interference is *Pcna*, which is necessary for DNA polymerase action and appears also to be upregulated in infection.

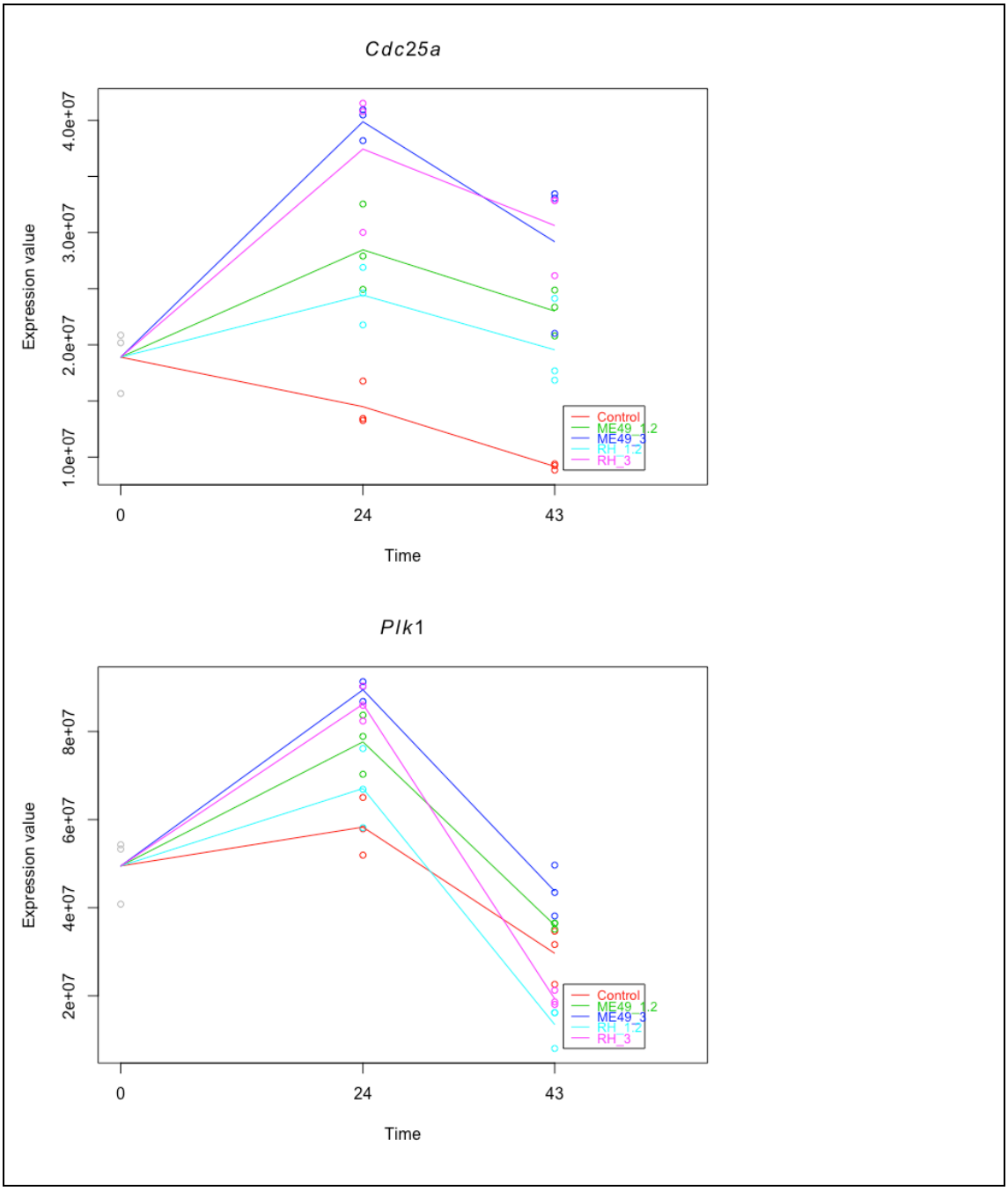


Figure 6.18: Expression Profiles of genes dysregulated in the uninfected sample and cell cycle-related genes.

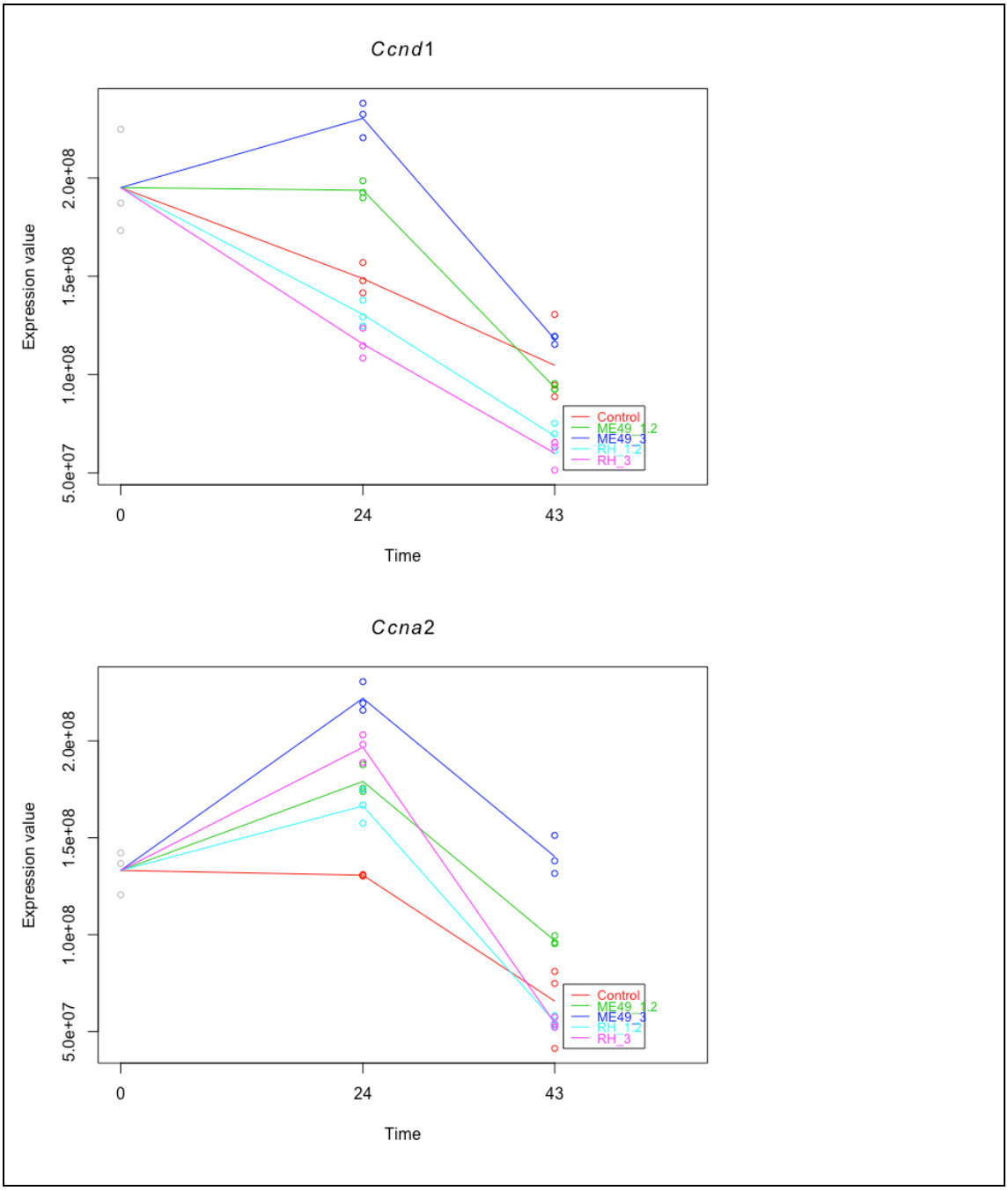
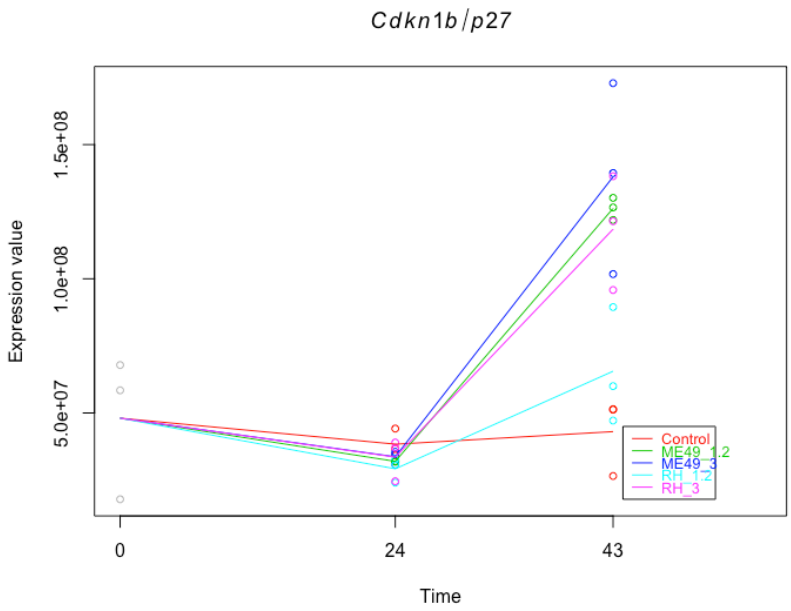


Figure 6.18, continued: Expression Profiles of genes dysregulated in the uninfected sample and cell cycle-related genes.



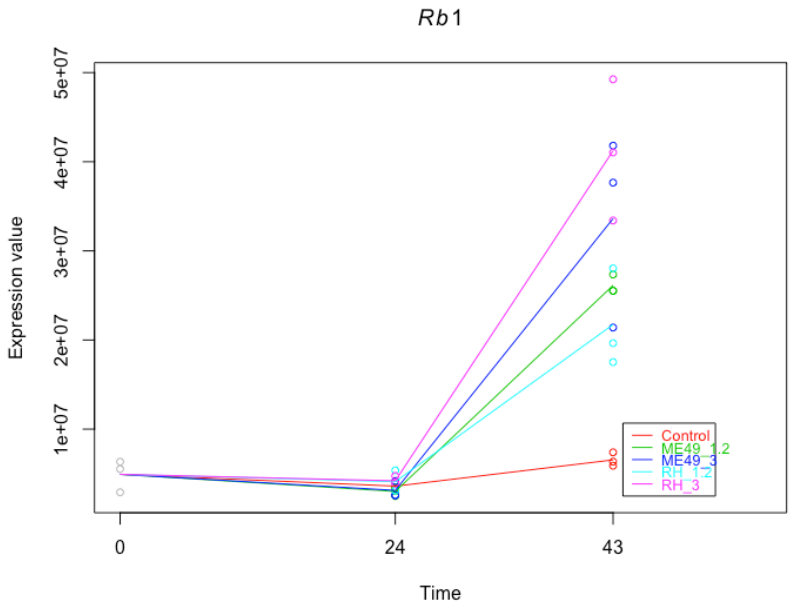


Figure 6.18, continued: Expression Profiles of genes dysregulated in the uninfected sample and cell cycle-related genes.

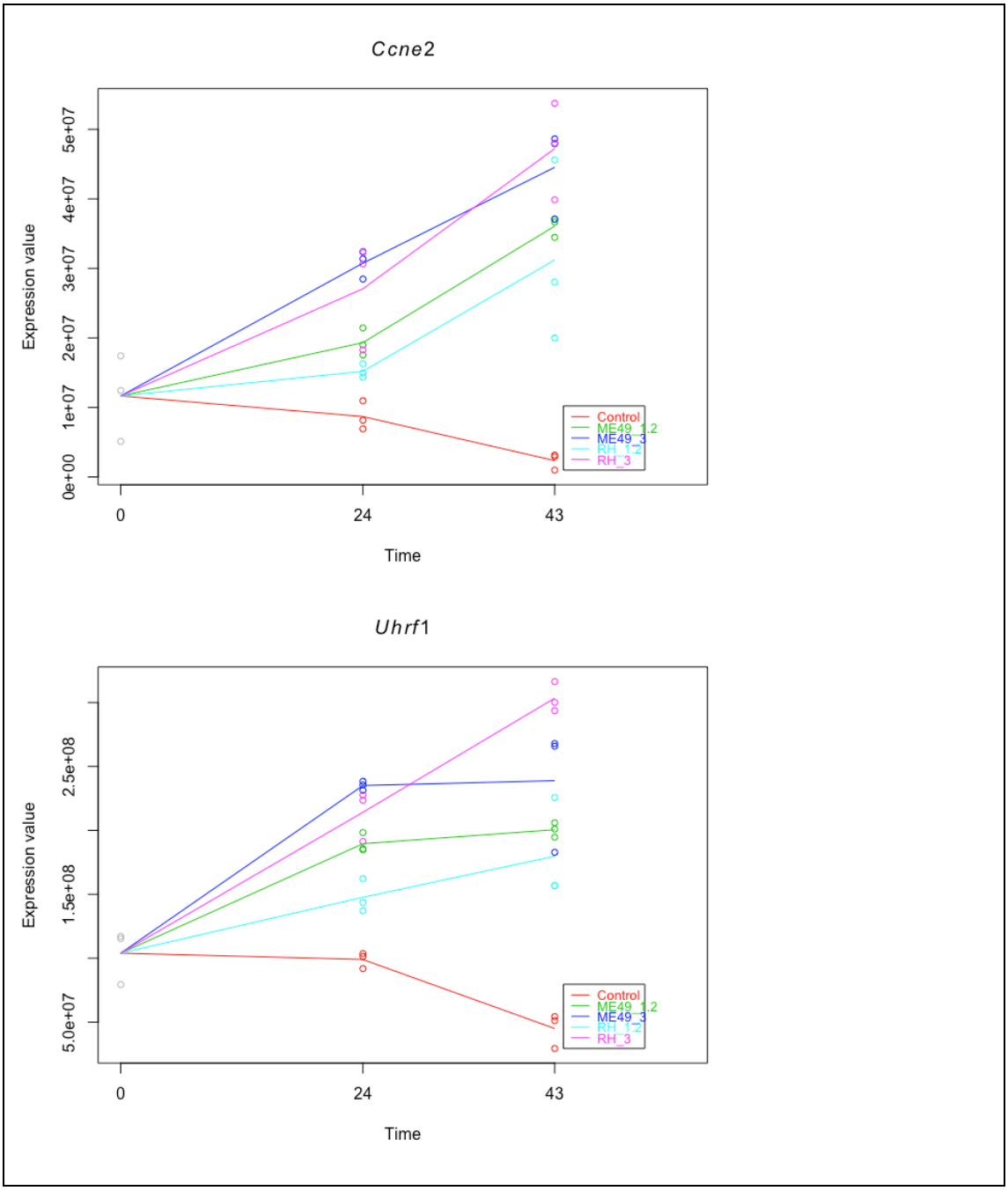


Figure 6.18: Expression Profiles of genes dysregulated in the uninfected sample and cell cycle-related genes.

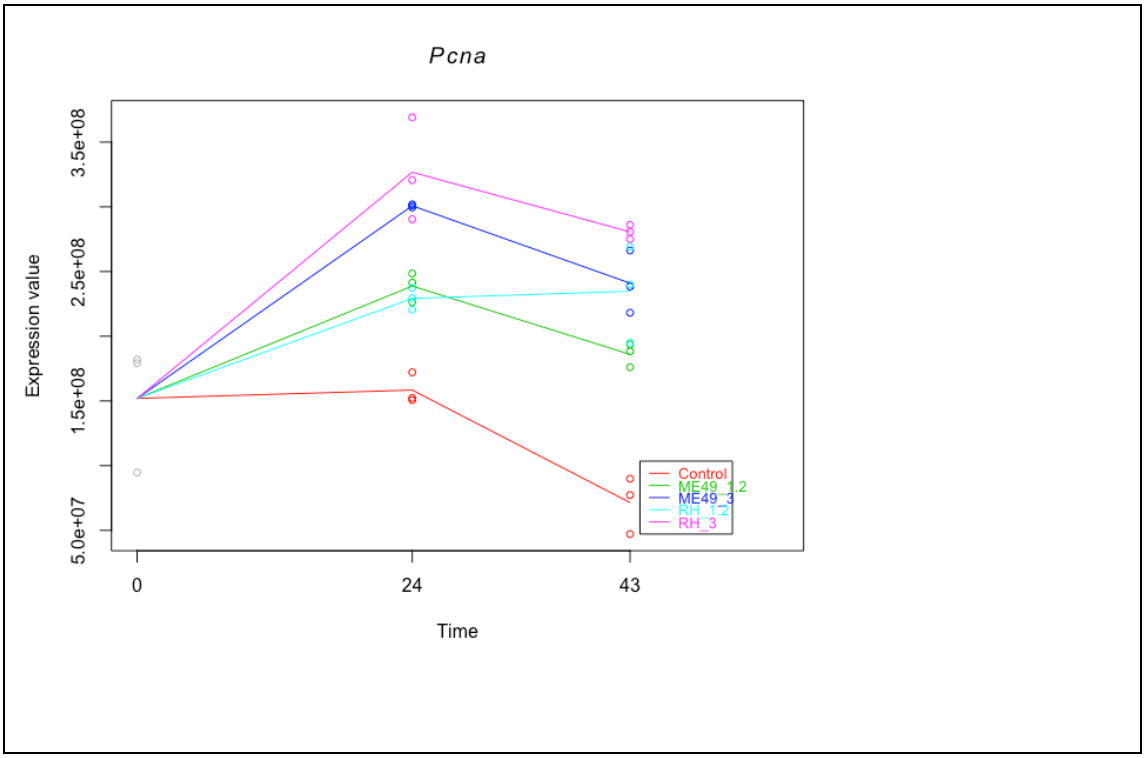


Figure 6.18, continued: Expression Profiles of genes dysregulated in the uninfected sample and cell cycle-related genes.

Hypoxia

It has been observed that *T. gondii* is able to stabilise HIF1A and that in fact HIF1A is necessary for proper rates of intracellular parasite growth (144). Hypoxia induced factor 1, alpha subunit is the component of the dimer that is oxygen-regulated.

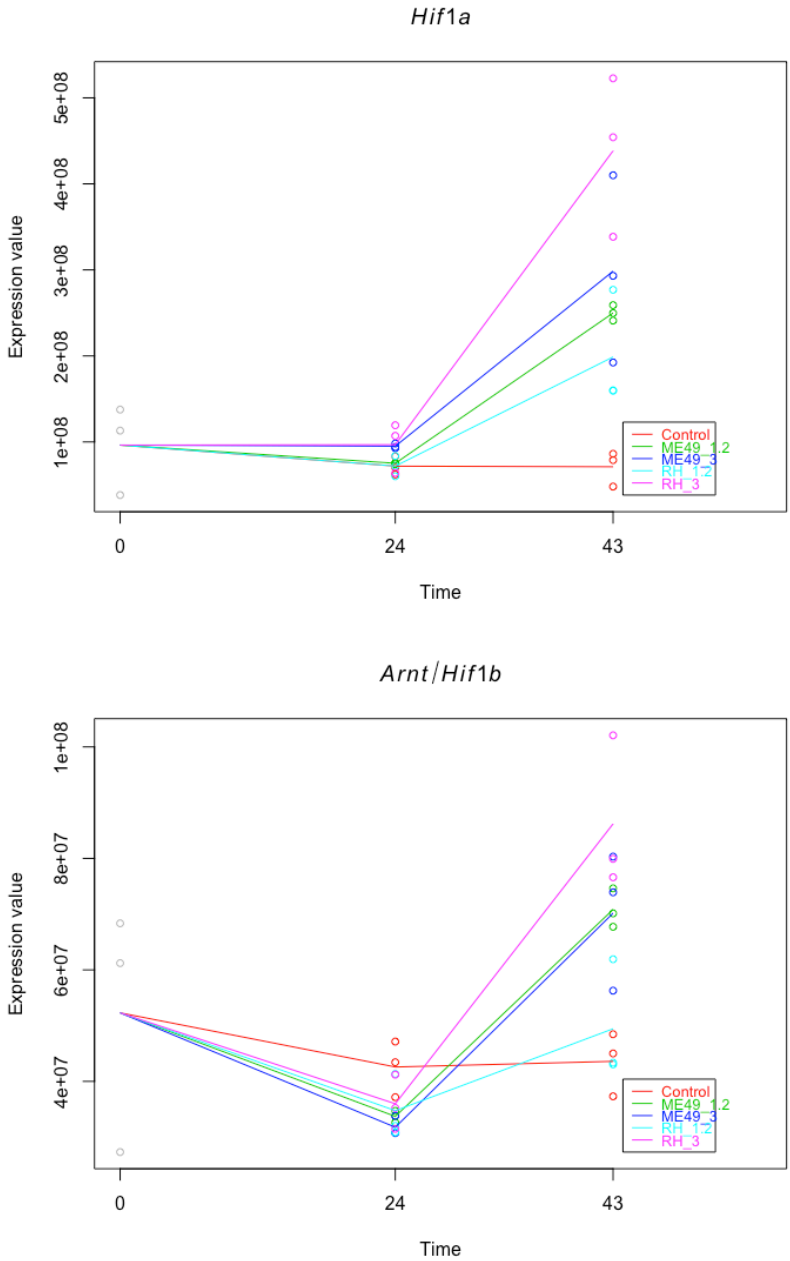


Figure 6.19. Expression profiles *of Hif1a* and *Arnt/Hif1b.*

Glycolysis and Cancer-Related Genes

The most striking profiles of glycolysis-related genes that were differentially-regulated in the infection timecourse were *Ldha* and *Hk2*, with both being highly upregulated as compared to both time and the uninfected control. These genes have been highly implicated in cancer processes – a metabolic pathway that emerged several times within the KEGG enrichment analyses.

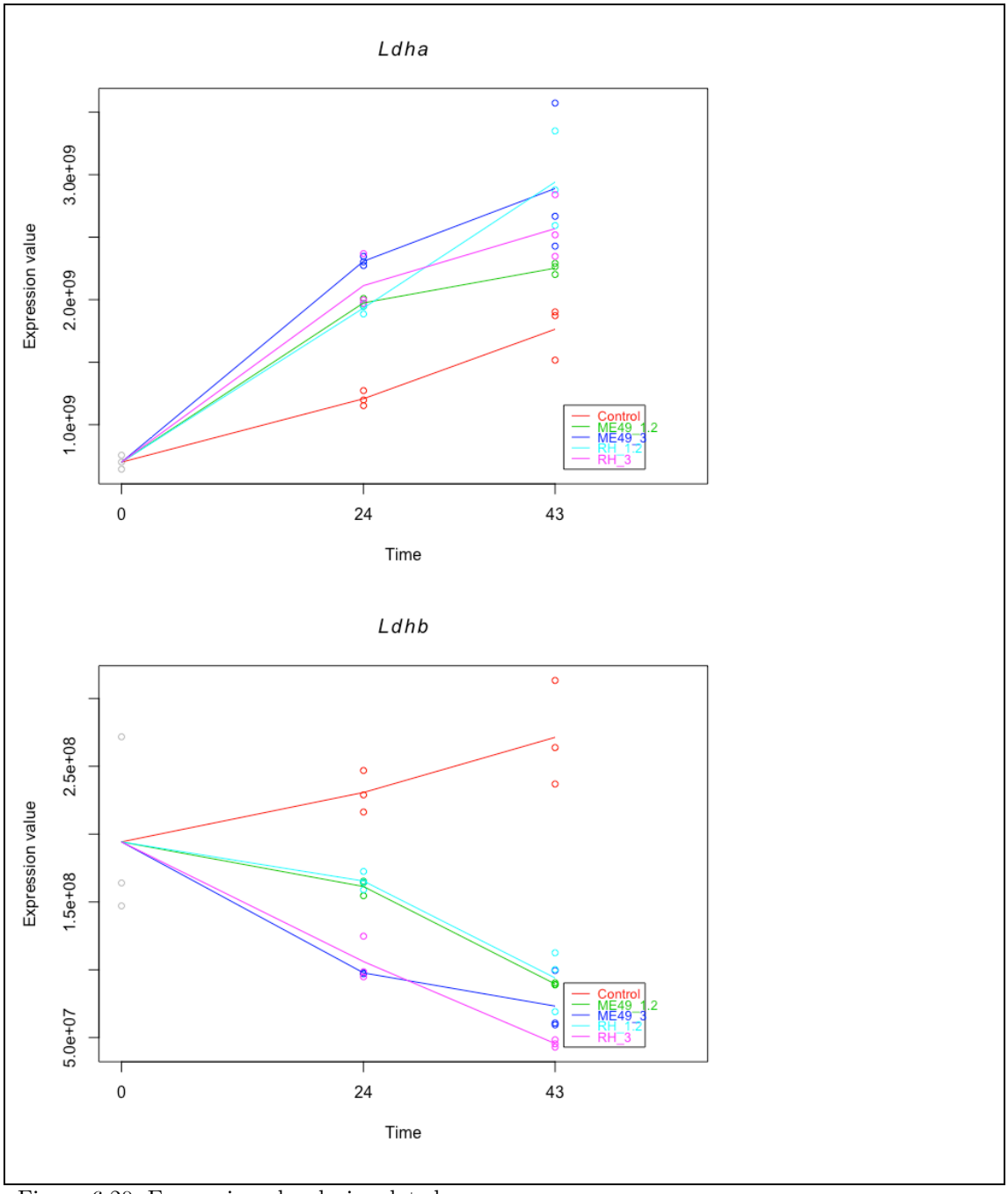
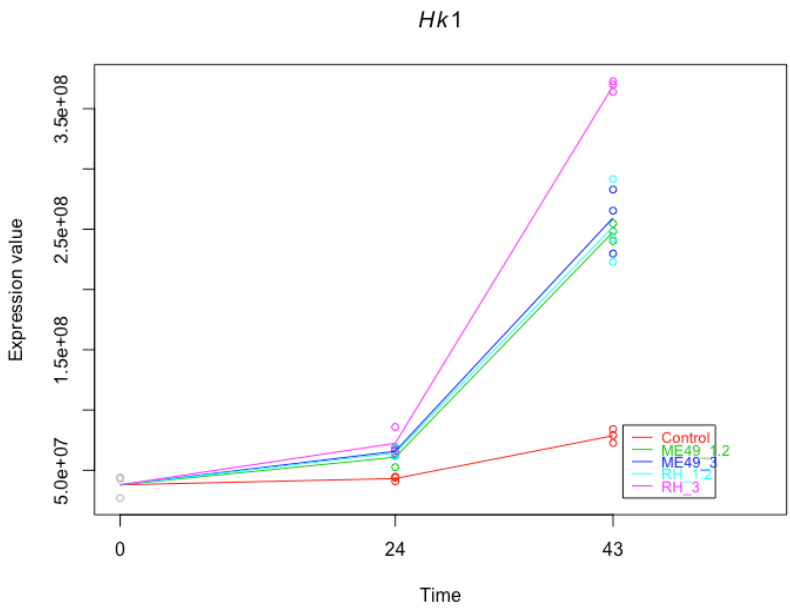


Figure 6.20. Expression glycolysis-related genes



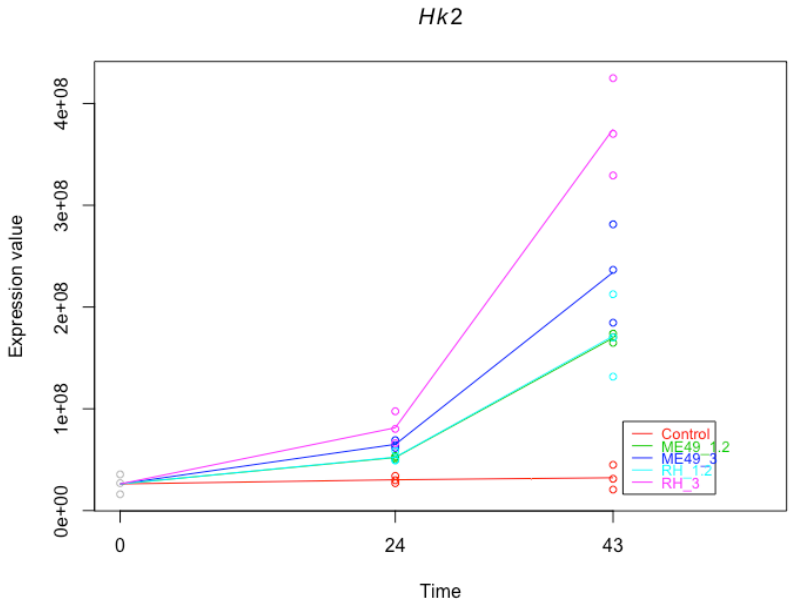


Figure 6.20, continued. Expression glycolysis-related genes

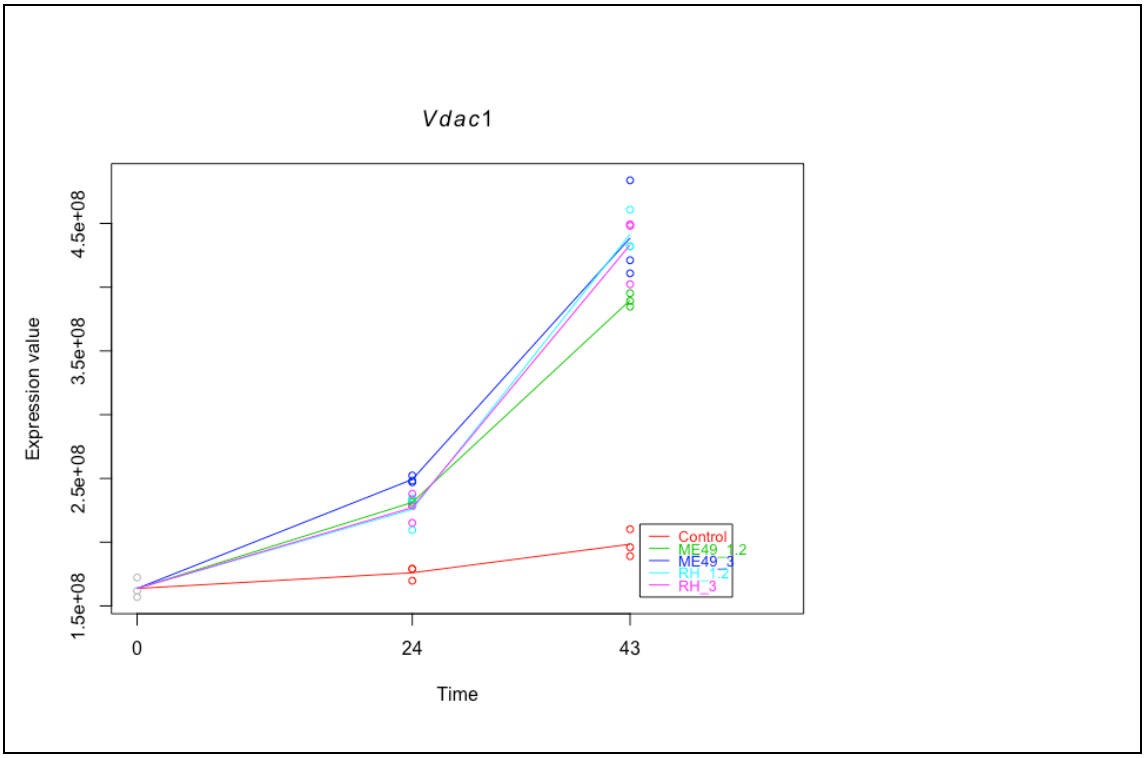


Figure 6.20, continued. Expression glycolysis-related genes

Vdac1 is thought to be part of the mechanism by which *Hk2* is able to remain mitochondrially-associated in highly-glycolytic conditions (198) and it also has potenti anti-apoptotic functions (199) and thus it is unsurprising that it should be elevated in all infection conditions. However, how these two mechanisms interact with each other is still unclear and is discussed further in 6.4.

Transporters

A number of transporters were shown to be differentially-regulated throughout the course of infection. Figure 6.12 shows the profiles of the lactate importer *Slc2a1* and *Mct1*, a lactate exporter. All strains clearly show an elevation of these transcripts, even from the 24hpi onwards.

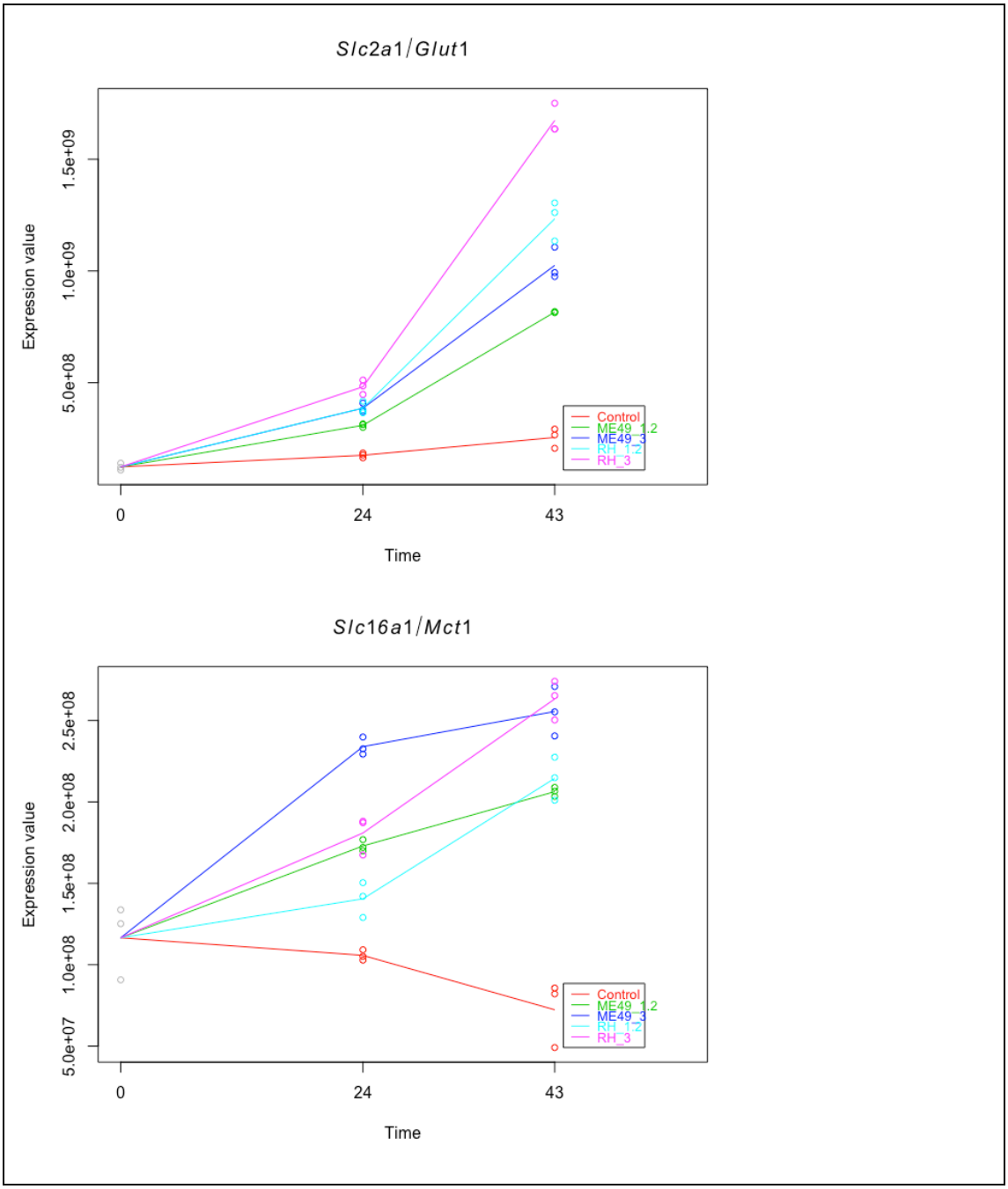


Figure 6.21. Expression genes encoding transporters

Sirtuins

While still relatively understudied, Sirtuins have been implicated in the control of processes that are known to be parasite-disrupted, such as apoptosis and glycolysis, perhaps through mediation of HK2. The role of *Sirtuins* in glycolysis and cancer is a controversial one, that is examined in **6.4**. Few generalisations emerge from their profiles (Figure 6.22).

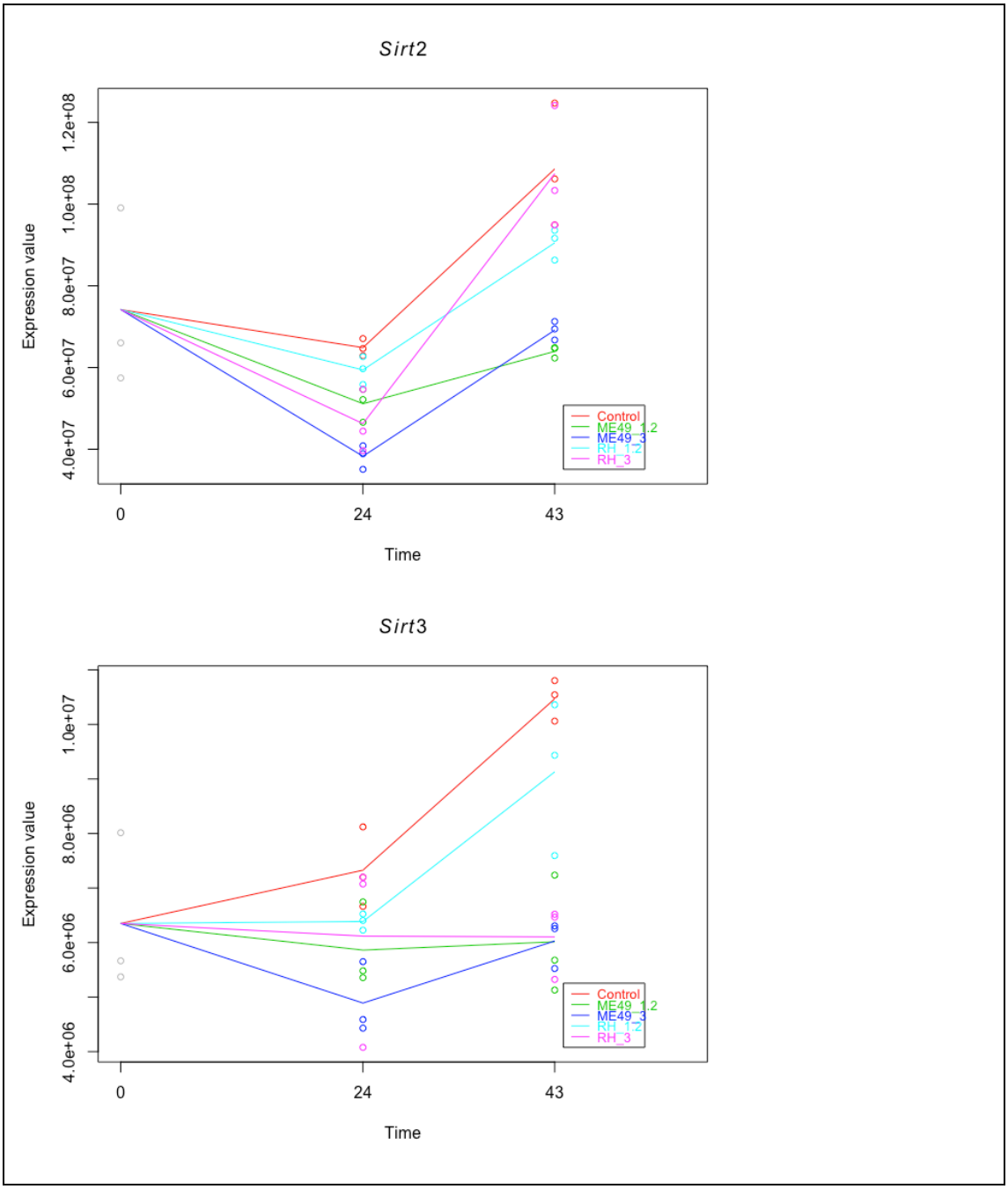


Figure 6.22. Expression profiles of sirtuin and sirtuin-related genes

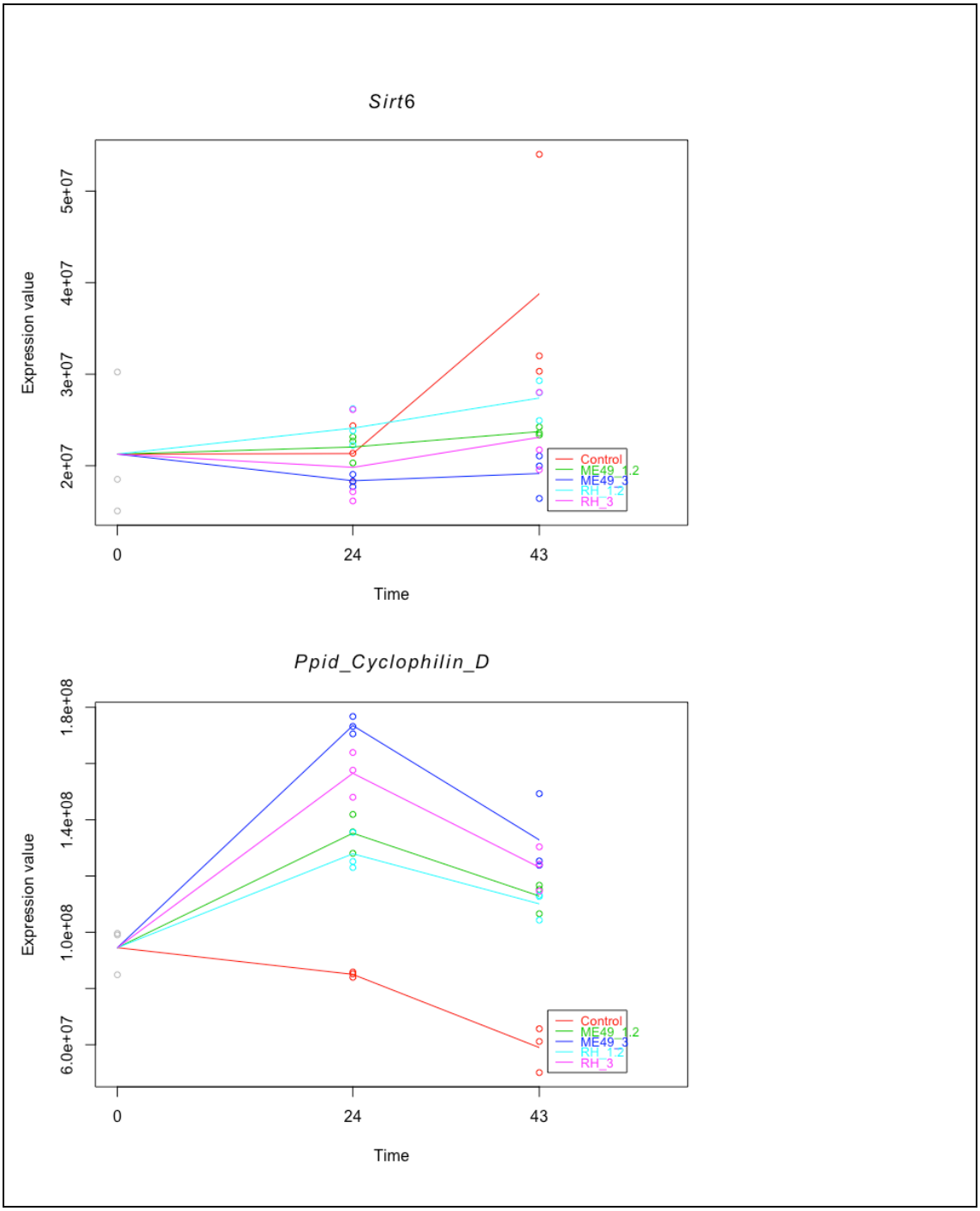


Figure 6.22, continued. Expression profiles of sirtuin and sirtuin-related genes

Glutamine-Related Genes

The repeated reference to "glutamine metabolism" in so many of the enrichment lists led me to look not just at this phenomenon but other aspects of cellular glutamine function. *Gls*, for instance encodes Glutaminase, which is seen to be upregulated in all the infected samples. This is also the case for

glutamine transporters *Slc7a5* and *Asct2*, both of which have been implicated in increased glycolysis in cancer cells.

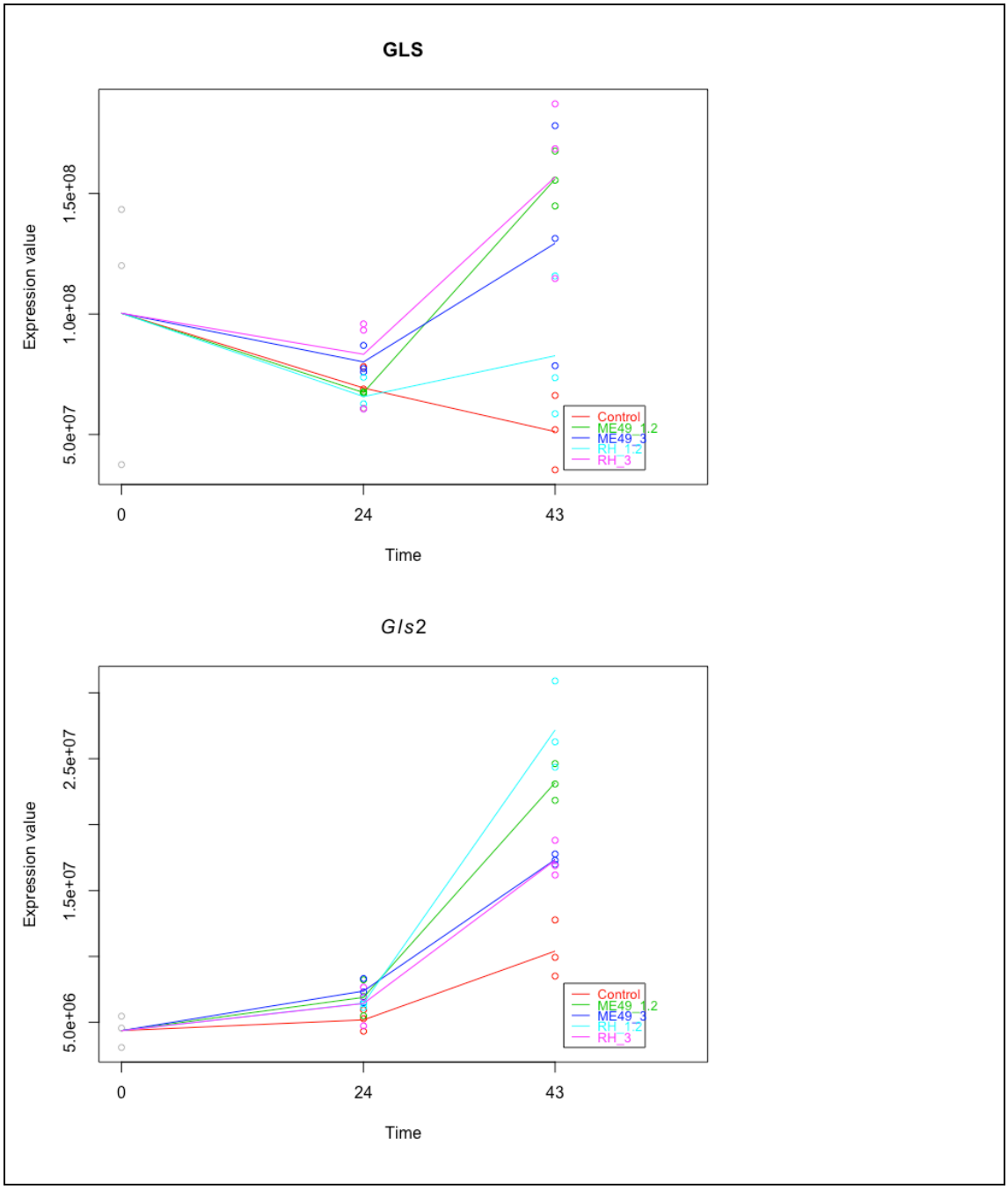


Figure 6.23. Expression profiles of genes with functions related to glutamine and glutaminolysis

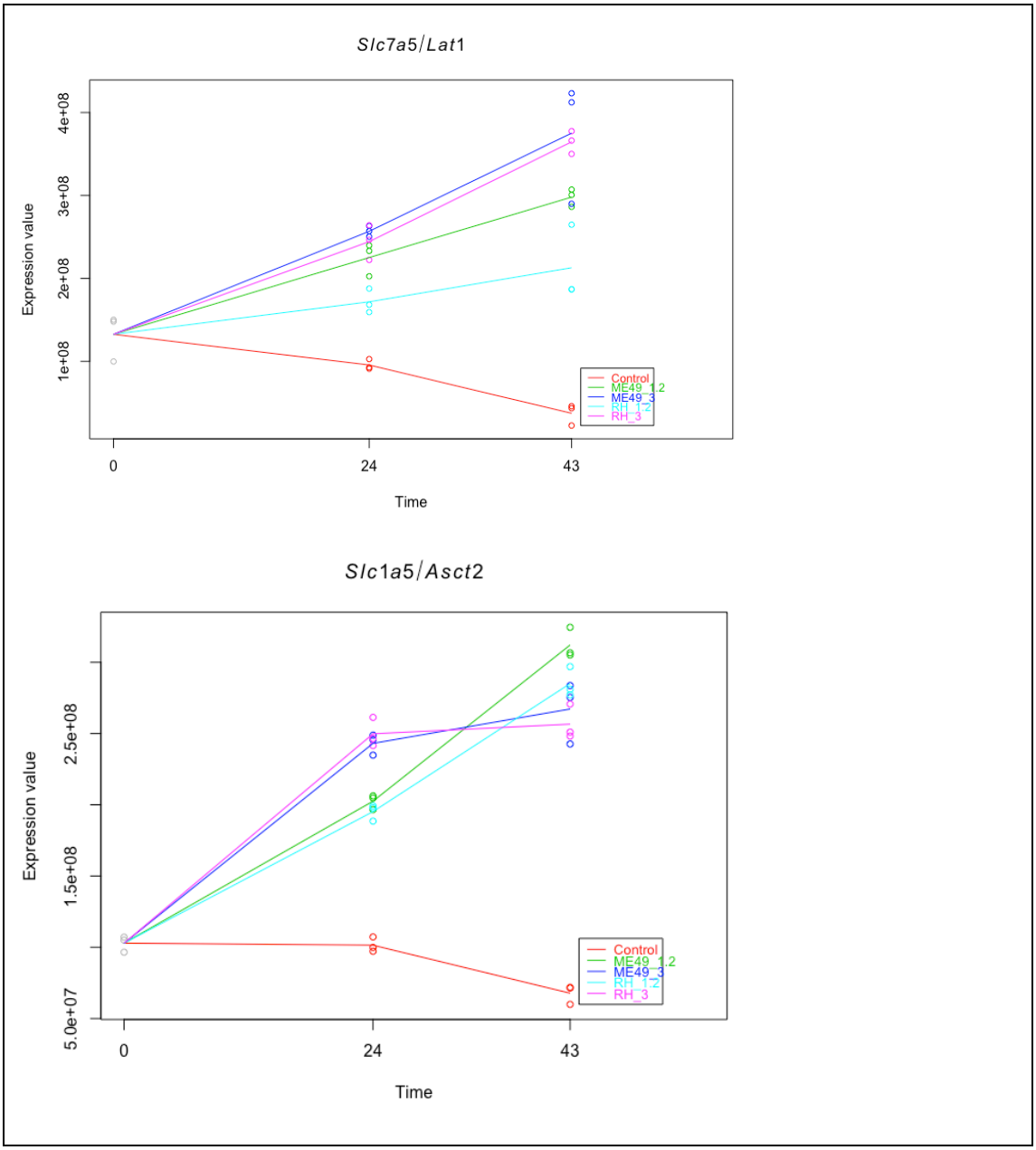


Figure 6.23, continued. Expression profiles of genes with functions related to glutamine and glutaminolysis

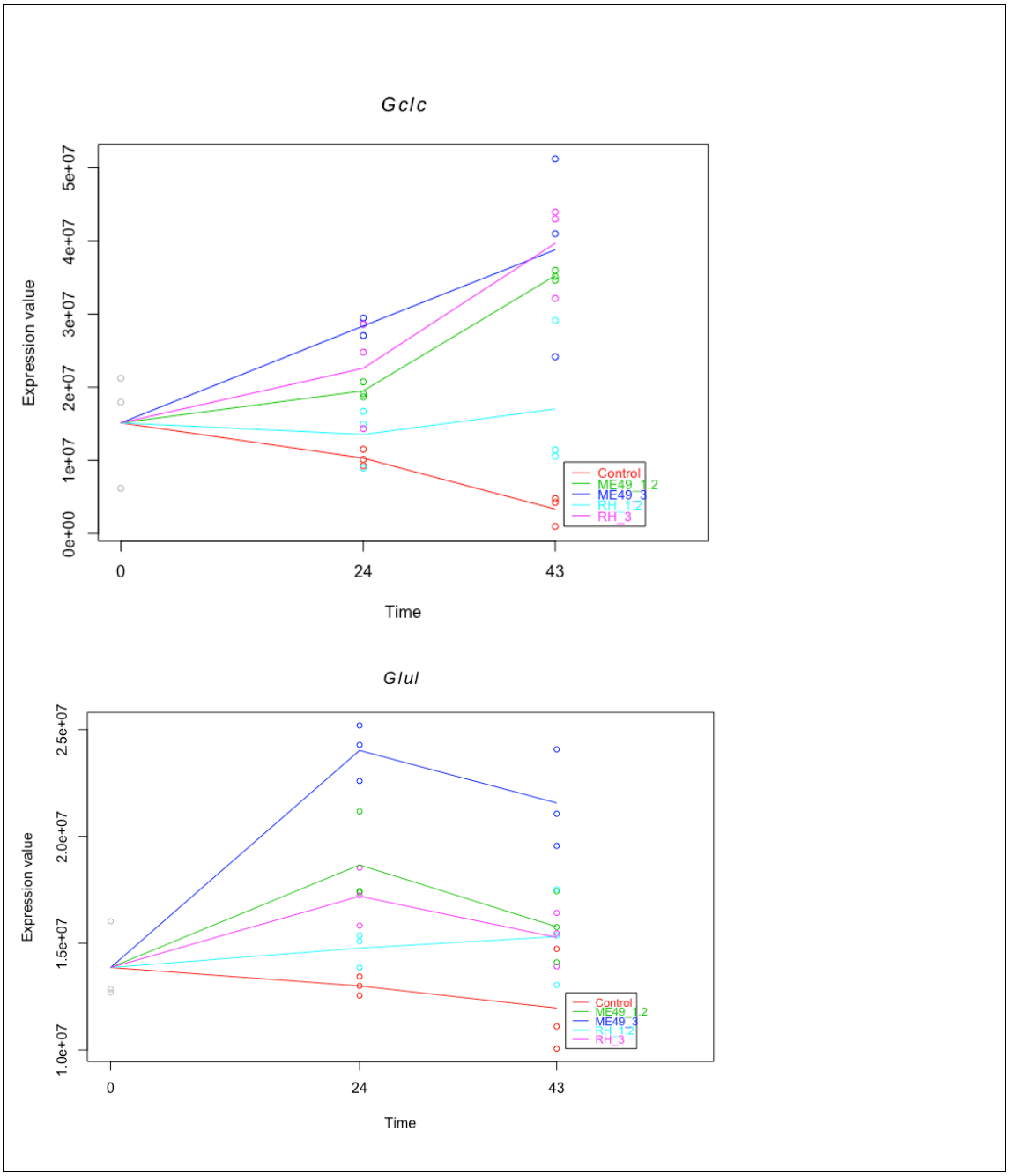


Figure 6.23, continued. Expression profiles of genes with functions related to glutamine and glutaminolysis

Glutathione and ROS-related

Many of the genes to do with the "glutamine metabolism" pathways had to do with the replenishment of glutathione – a potent protector against ROS. These include from uncoupling proteins (*Ucp2*), transporters (*Slc7a11*) and metabolic enzymes (*Idh1*).

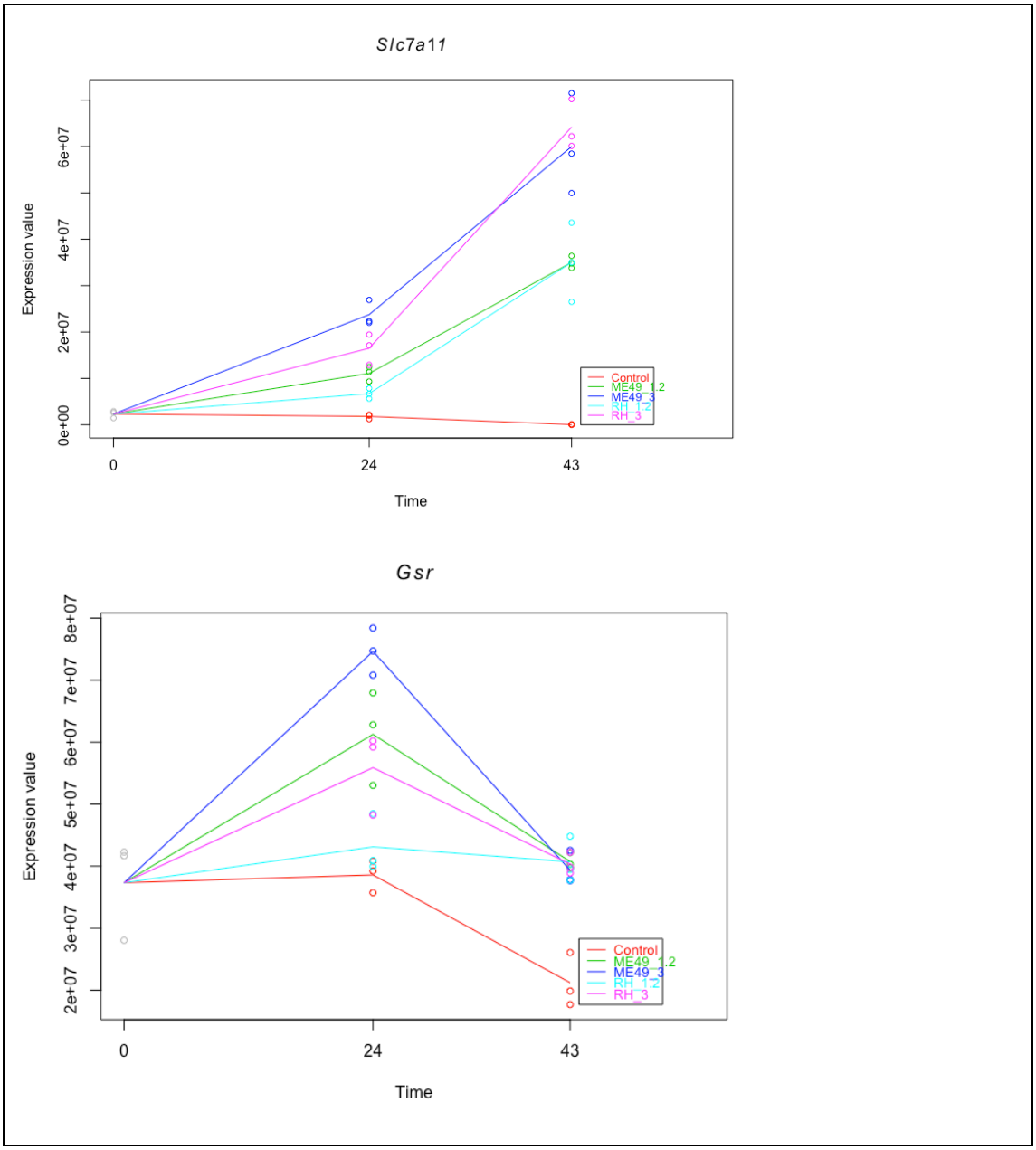


Figure 6.24. Expression profiles of genes with functions related to glutathione and ROS

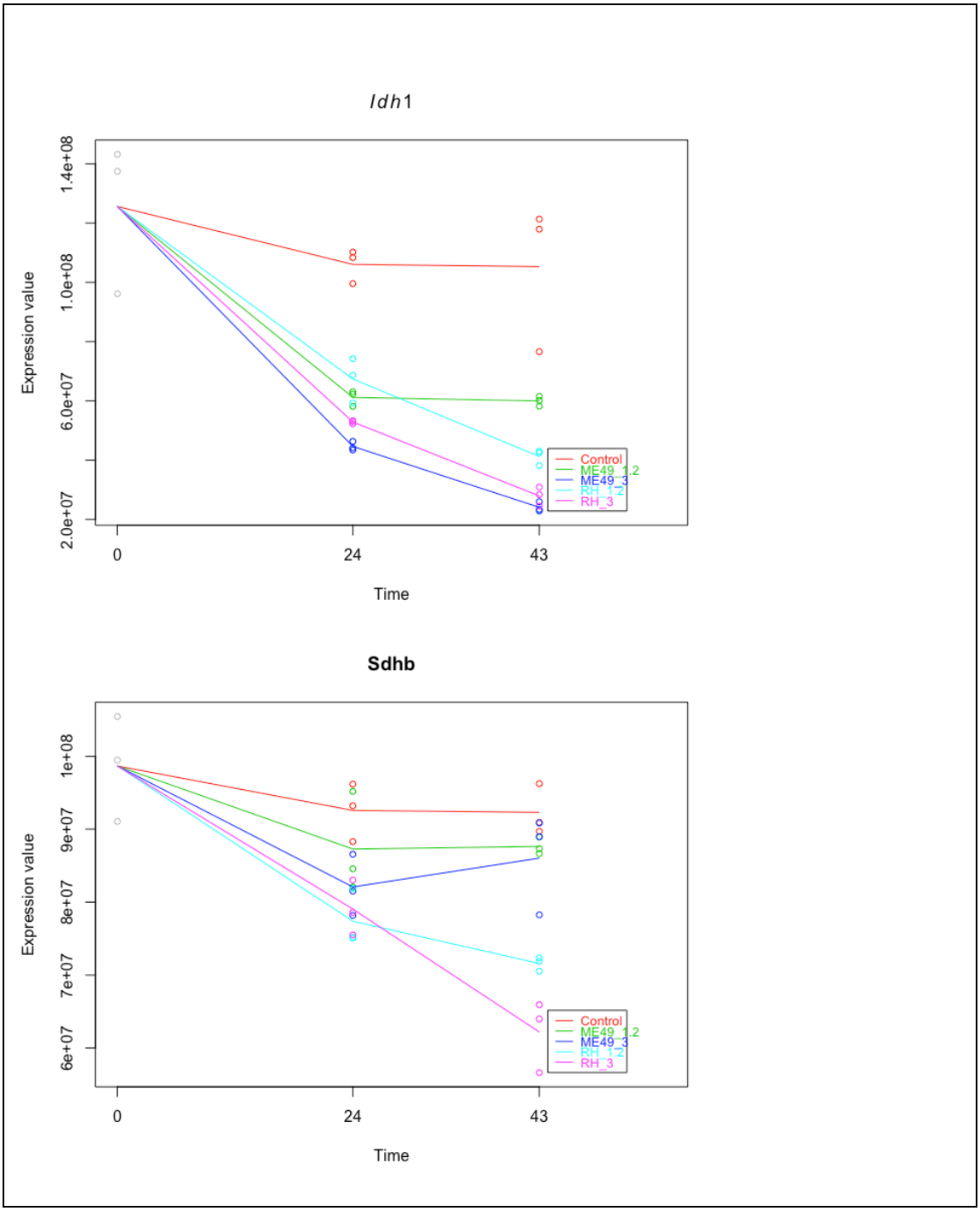


Figure 6.24, continued. Expression profiles of genes with functions related to glutathione and ROS

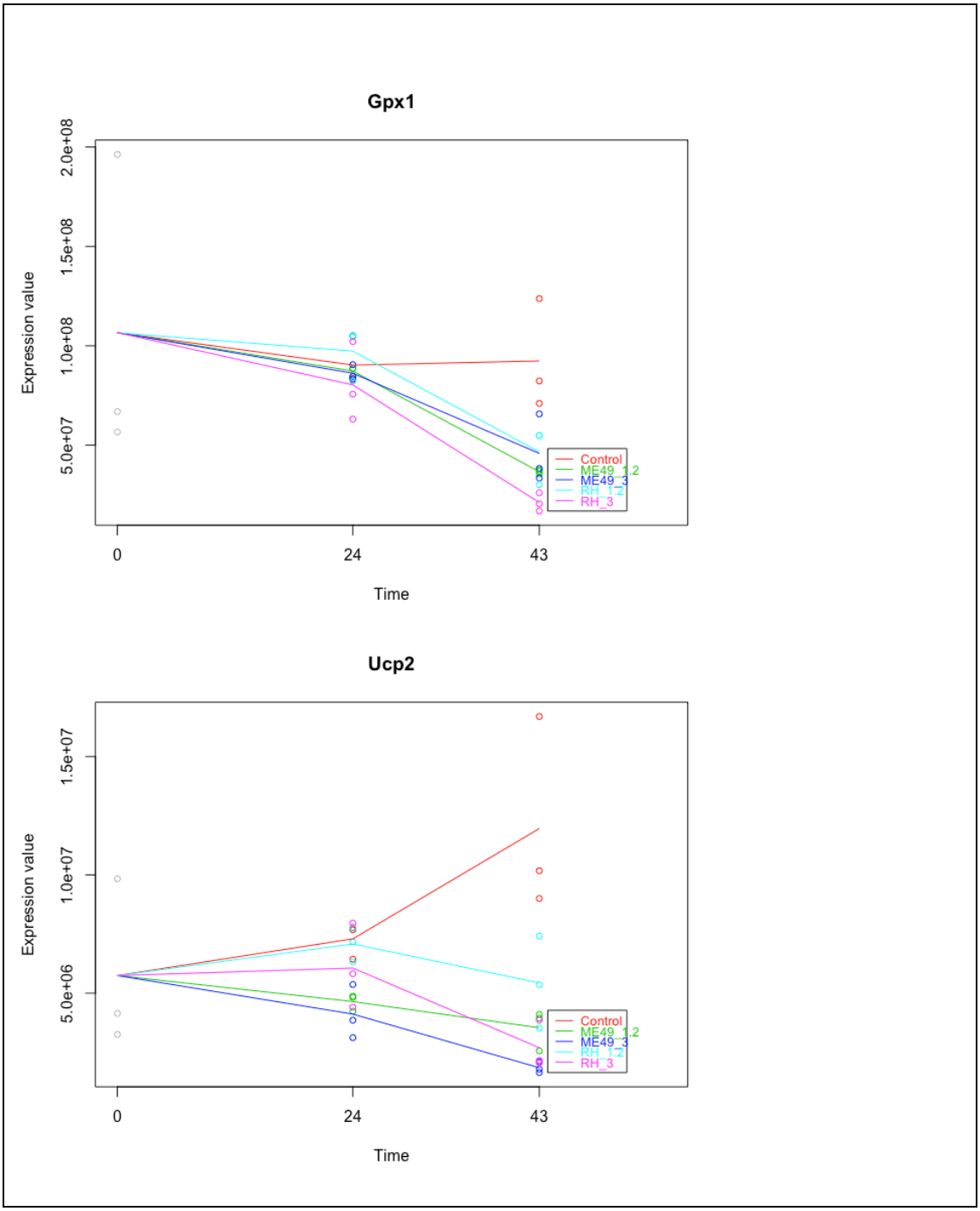


Figure 6.24. Expression profiles of genes with functions related to glutathione and ROS

Nucleotide Biosynthesis

Given *T. gondii* status as a scavenger and an auxotroph, I profiled several genes from nucleotide synthesis pathways, especially since this had emerged as a significanltly-enriched KEGG pathway as well.

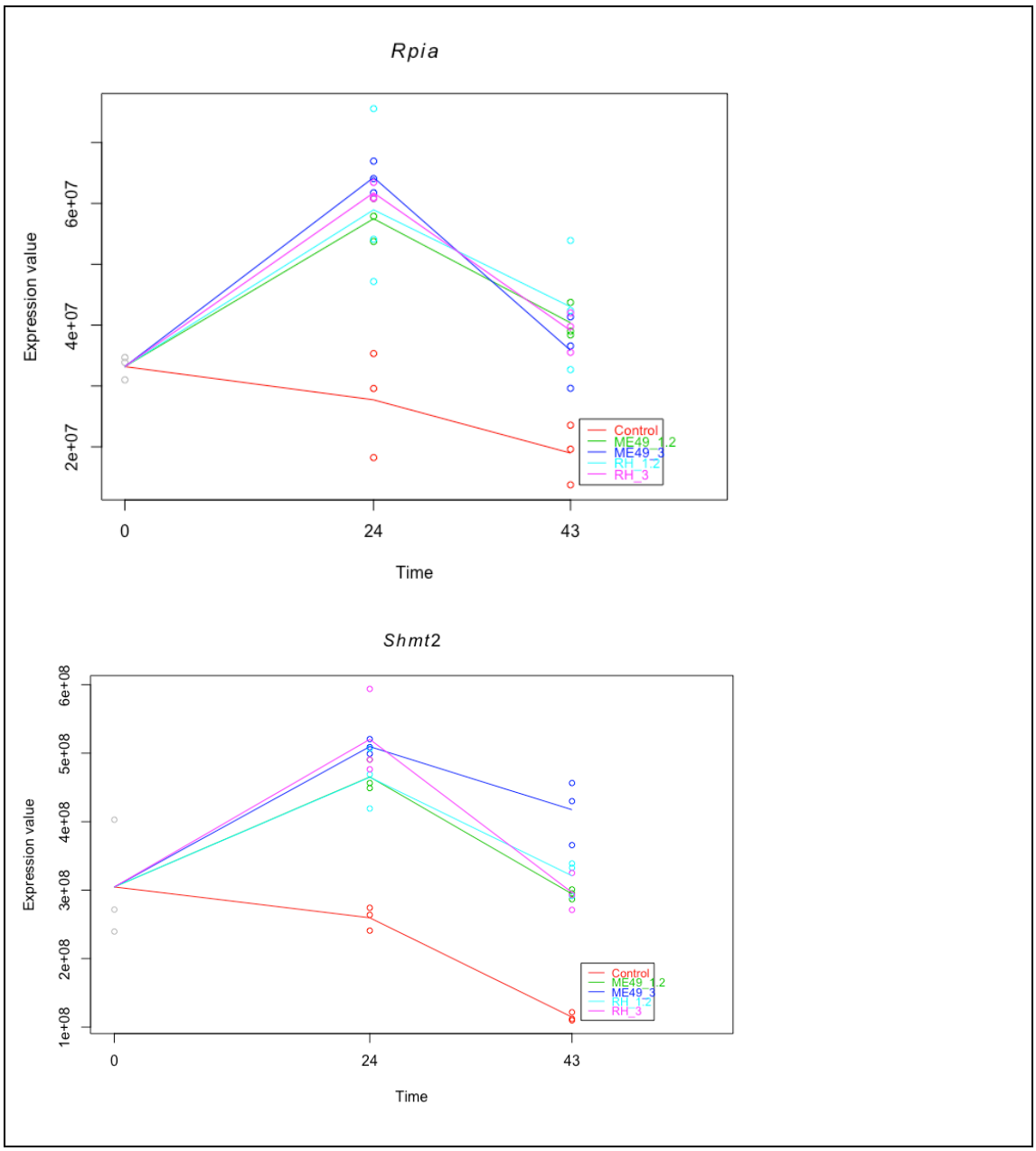


Figure 6.25. Expression profiles of genes related to nucleotide biosynthesis

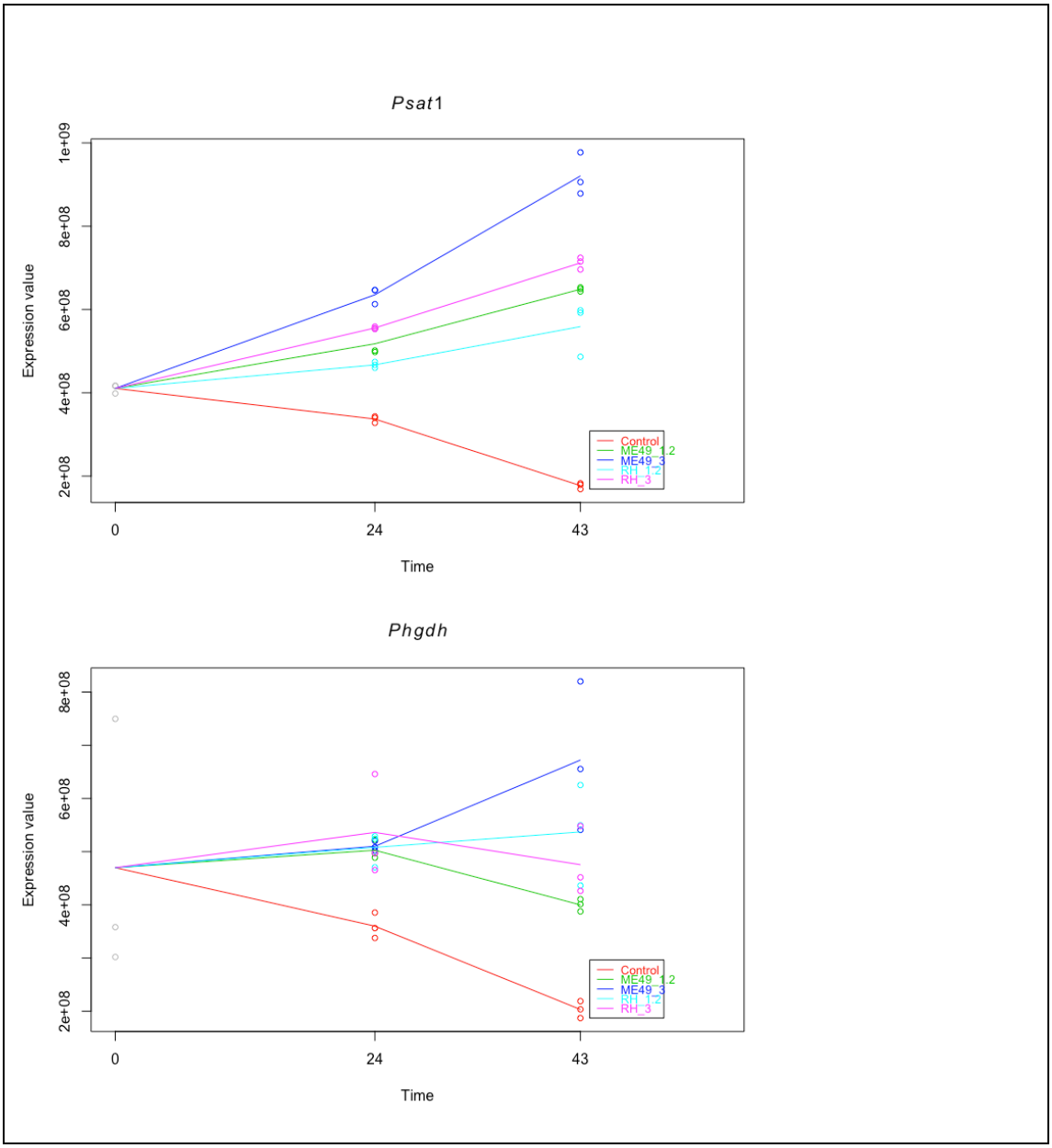


Figure 6.25, continued. Expression profiles of genes related to nucleotide biosynthesis

Fatty Acid Synthesis

T. gondii is well-known to scavenge lipids from the host cell, including cholesterol (40) and palmitate (200). Host cell synthesis of these could require upregulation of the genes involved in fatty acid synthesis. The time-profiles of these genes is quite curious, with pyruvate dehydrogenase (*Pdha)* exhibiting a sustained elevation as compared to the control sample, whereas the gene for fatty acid synthase, *Fasn*, appears to peak earlier in infection. The two genes profiled for malonyl decarboxylation and transacylation (*Mlycd* and *Mcat*, respectively) both seem to be downregulated in all infection conditions as compared to the uninfected control.

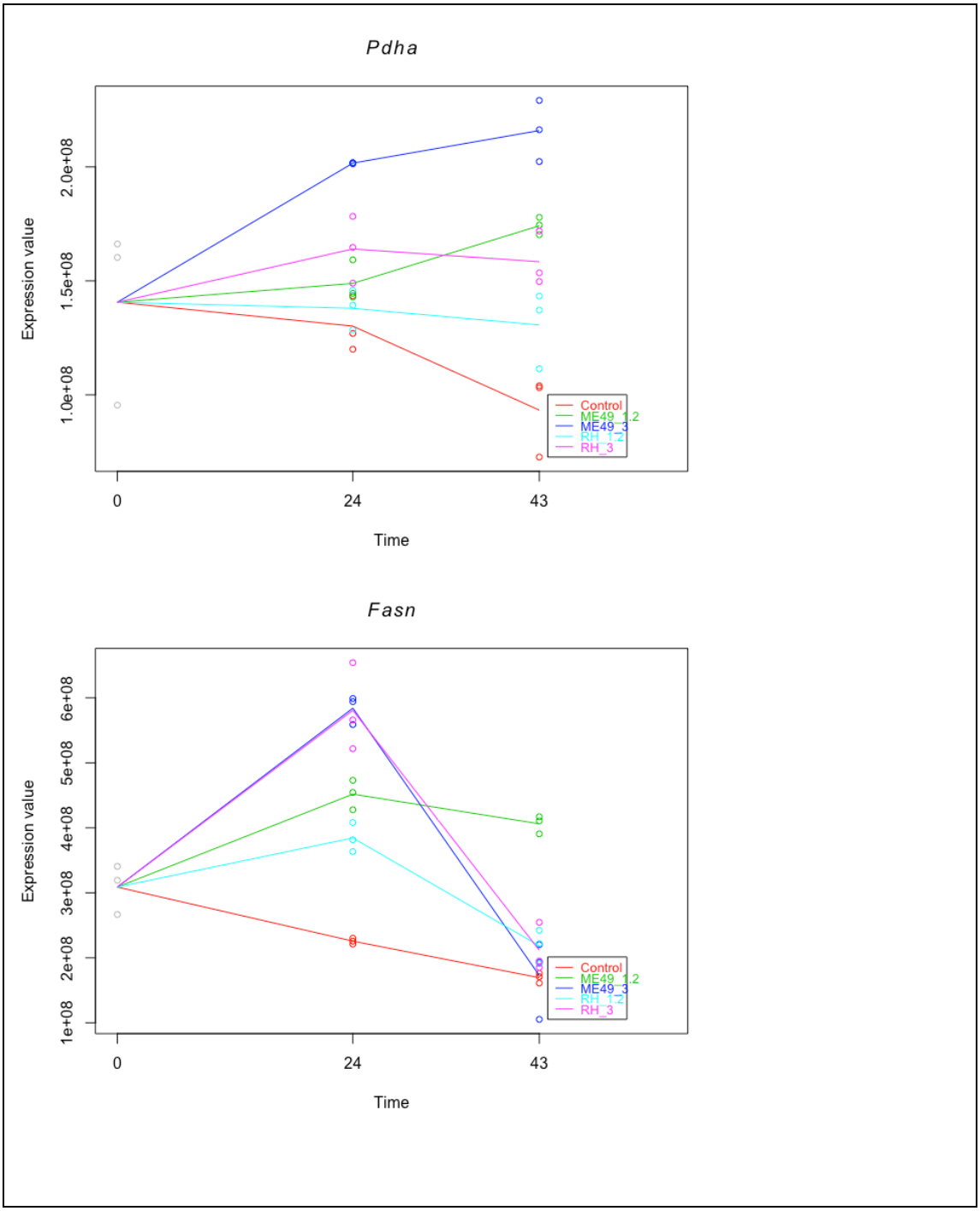


Figure 6.26. Expression profiles of genes related to fatty acid synthesis

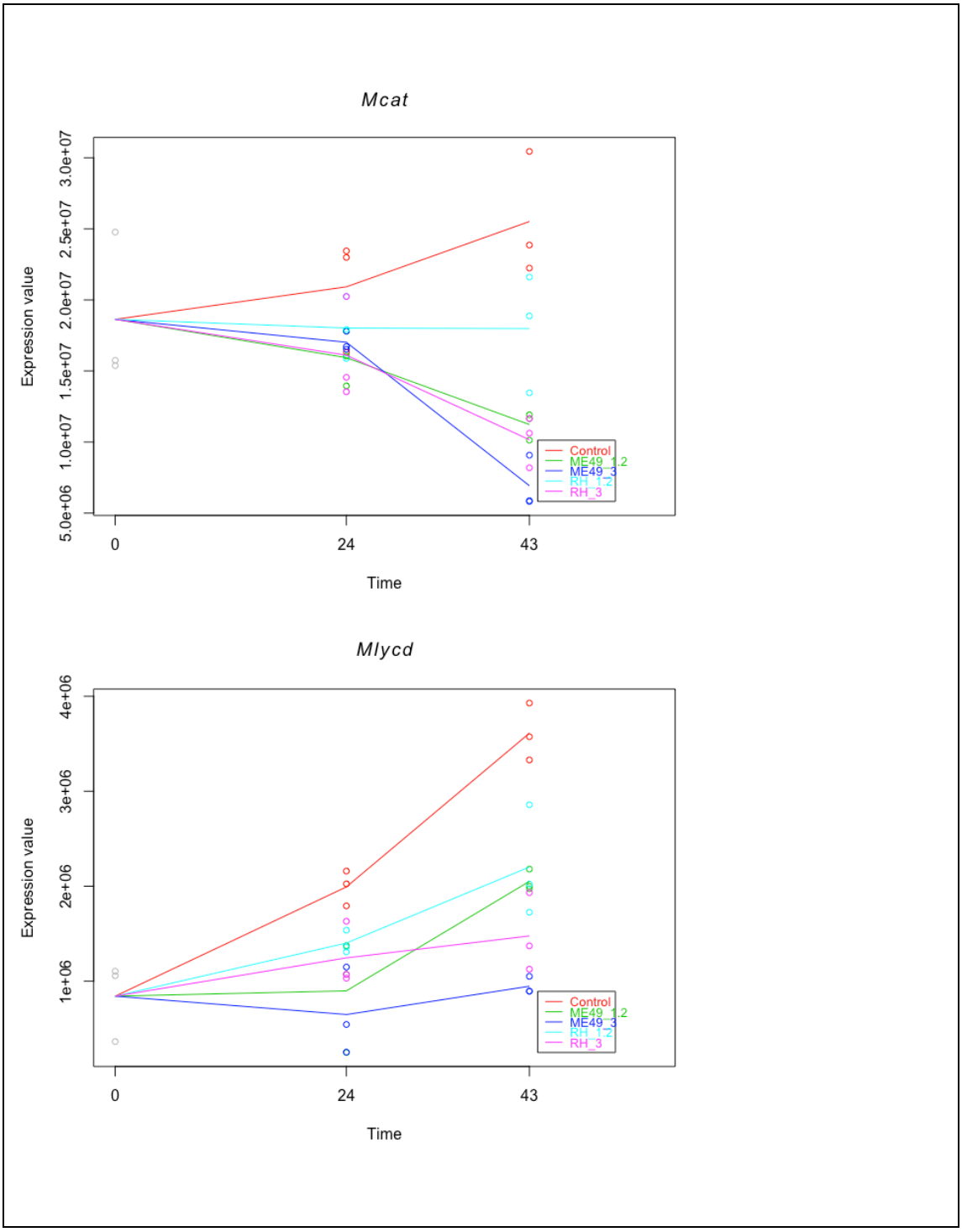


Figure 6.26, continued. Expression profiles of genes related to fatty acid synthesis

Myelocytomatosis oncogene (*Myc*) and *Myc*-related

This potent oncogenic transcription factor is known to be upregulated (translationally) by *T. gondii* infection (21). It's binding partner *Max* appears to be dysregulated in a strain-dependent manner, which is discussed in **6.4**. Other related genes, *Jnk* and *Jun* which both act to regulate *Myc* in contrast are upregulated in all infection conditions.

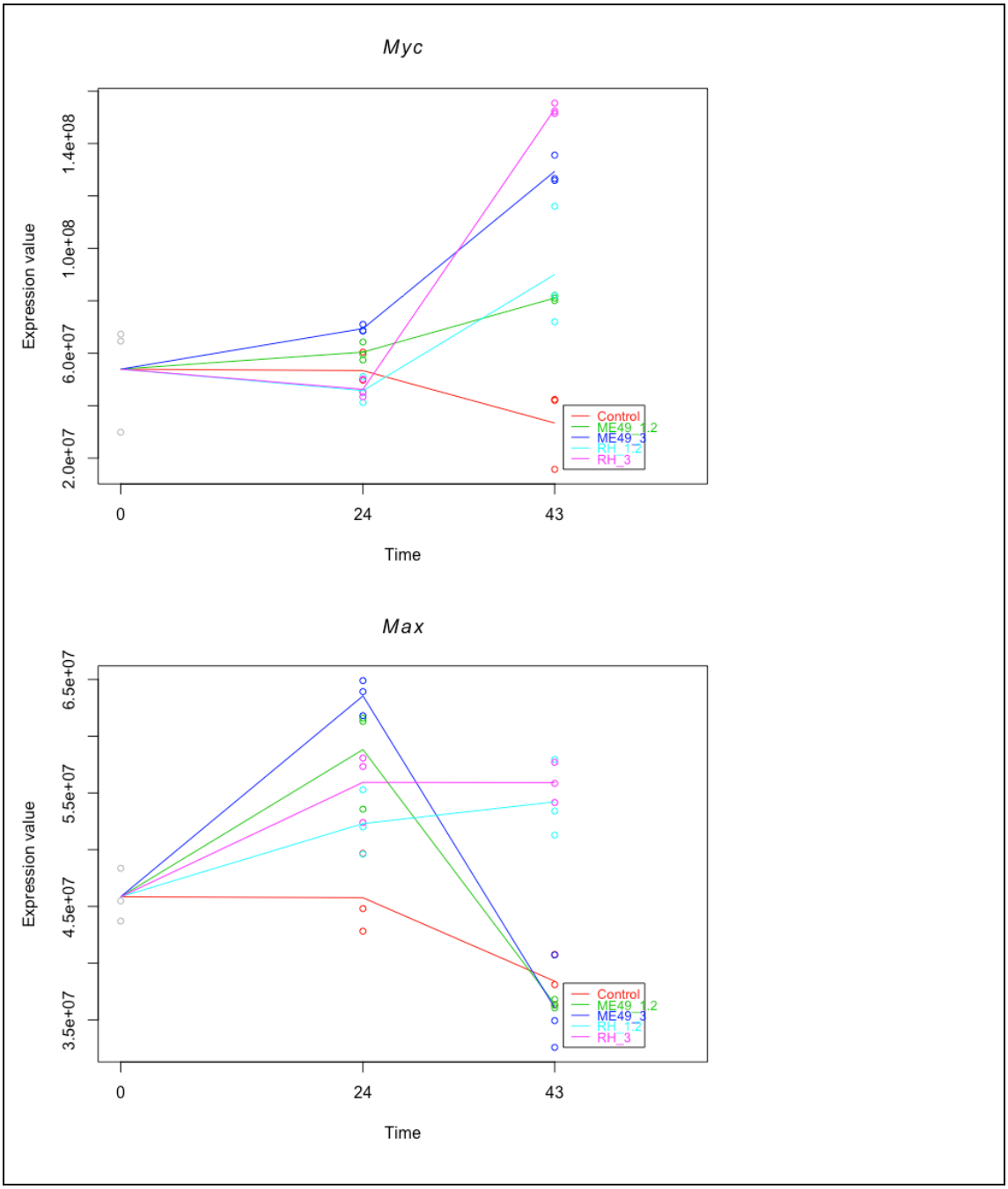


Figure 6.27. Expression profiles of *Myc* and *Myc*-related genes

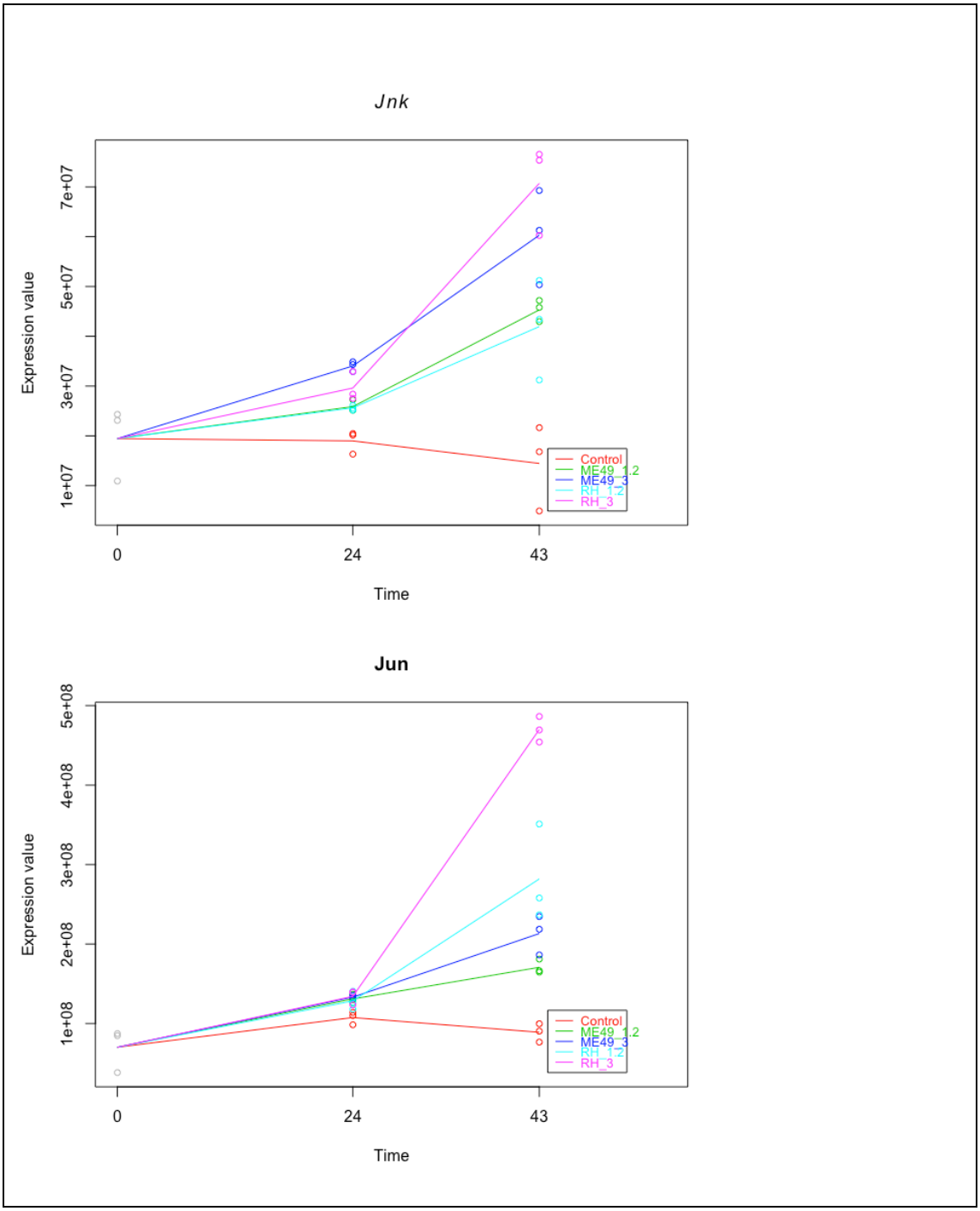


Figure 6.27, continued. Expression profiles of *Myc* and *Myc*-related genes

Apoptosis-Related Genes

Apoptosis is a process that the parasite is able to effectively evade, and it is mediated by both pro-apoptotic and anti-apoptotic factors. Apoptosis and the mechanisms by which *T. gondii* modulates this pathway is discussed further in **6.4**, especially in relation to the NFkB family.

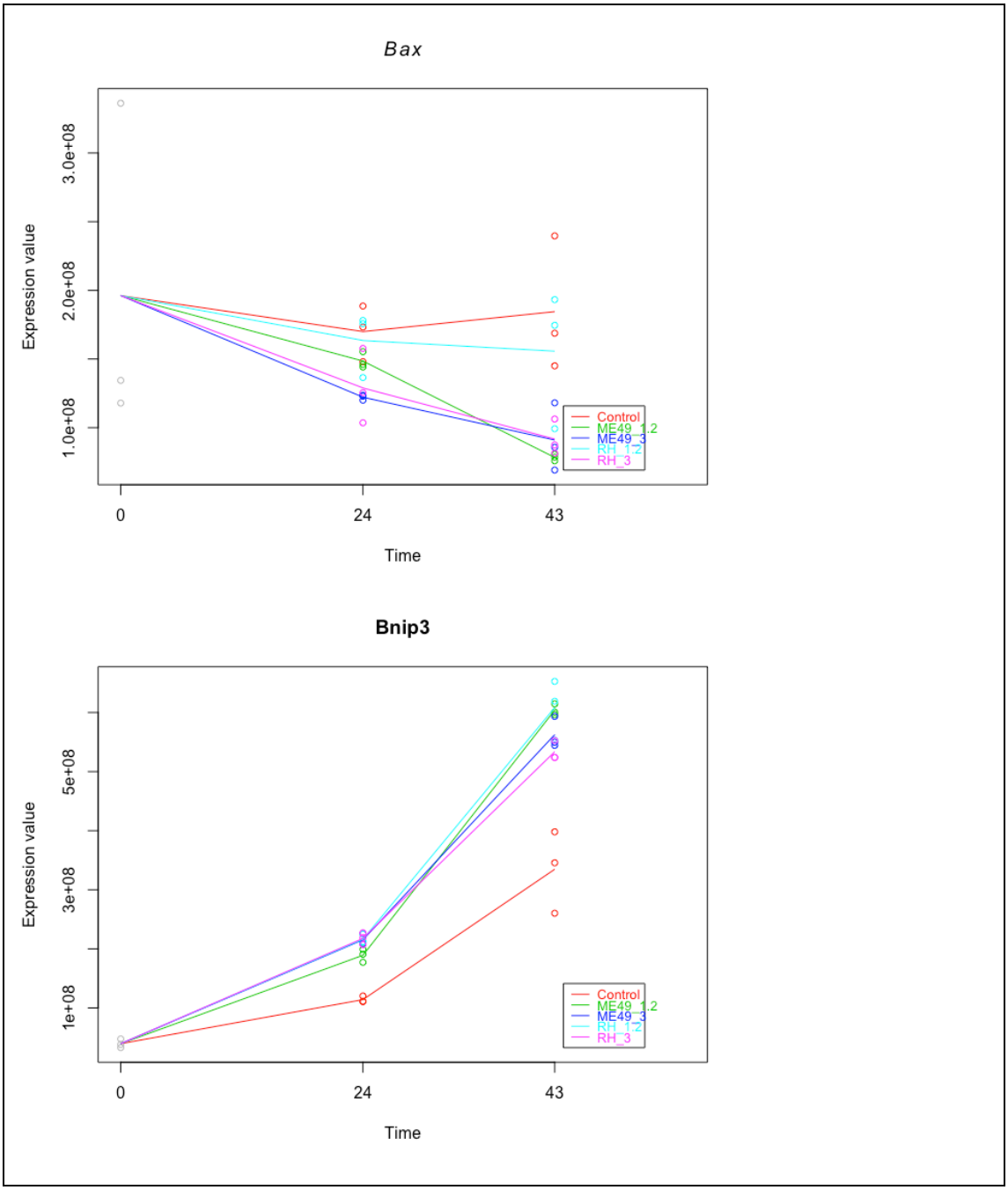


Figure 6.28, continued. Expression profiles of genes related to apoptosis

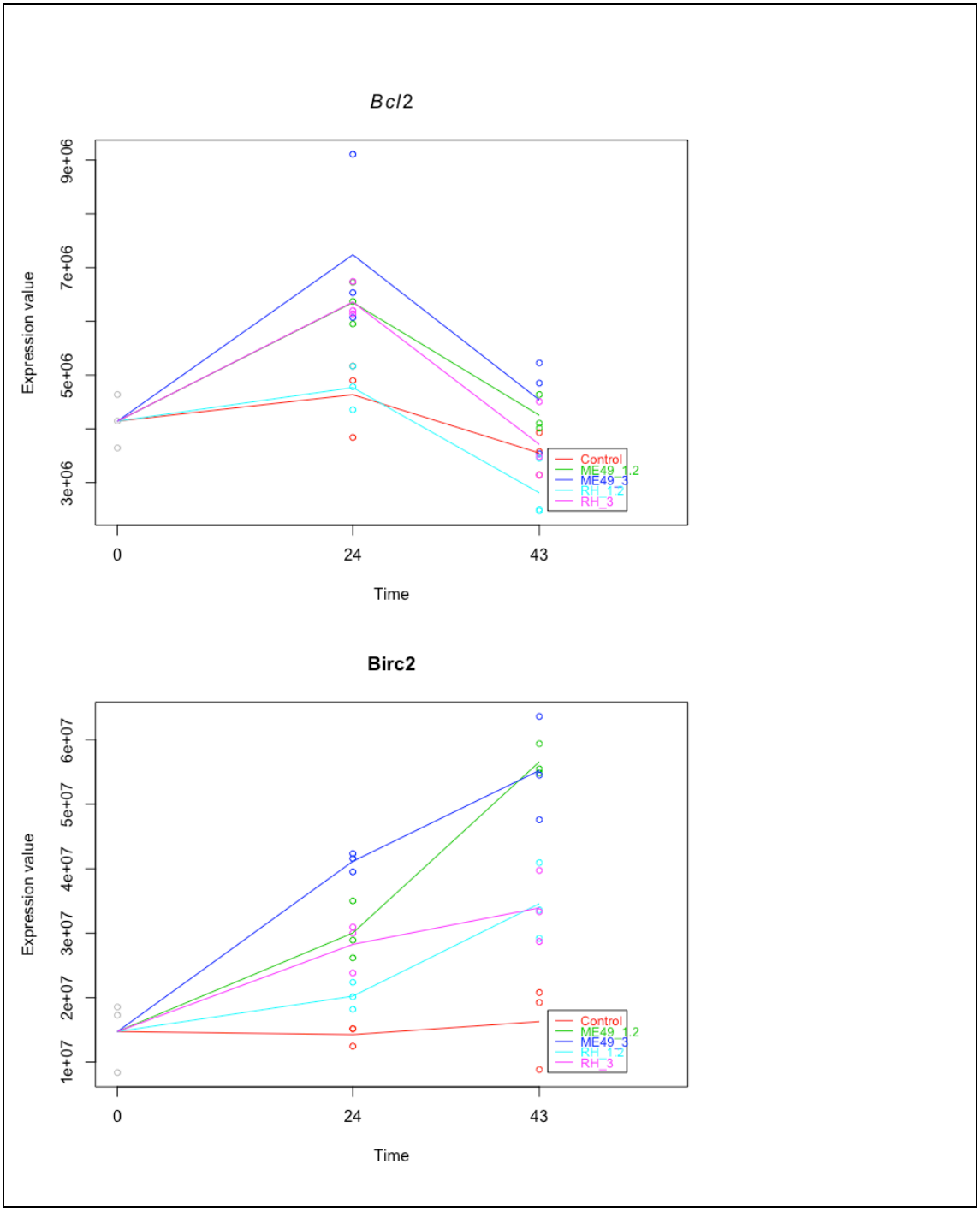


Figure 6.28, continued. Expression profiles of genes related to apoptosis

Transformation-related protein 53 (*Trp53*) and *Trp*-related

Akin to MYC, TRP53 is a potent transcription factor in the realm of cancer, except that it is an tumour suppressor. There are also numerous reported interactions with the *Myc*. Given the current lack of clarity surrounding *T. gondii*'s effects on this pathway, I looked at both *Trp53* expression and that of related regulators.

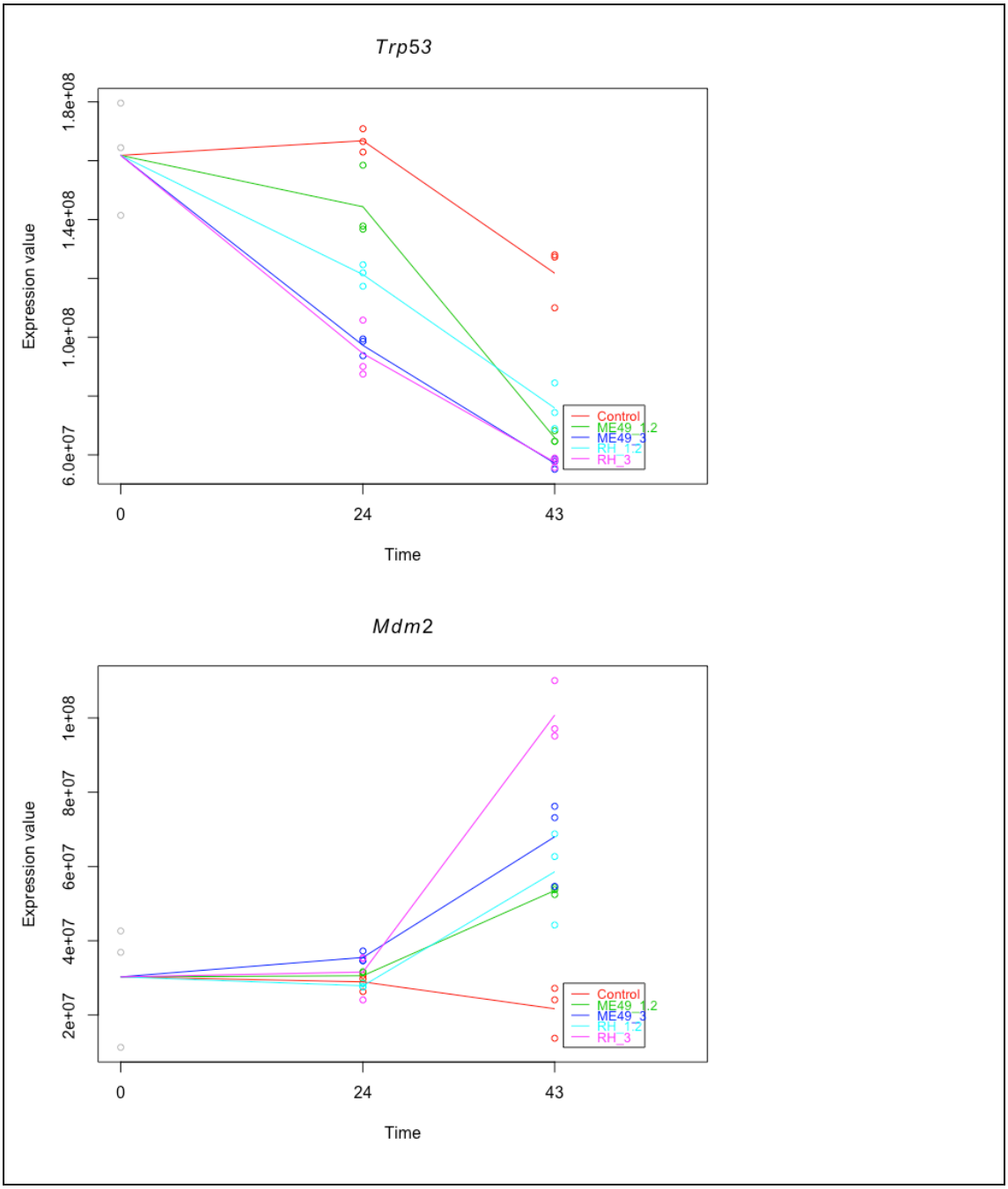


Figure 6.29. Expression profiles of genes related to *Trp53*.

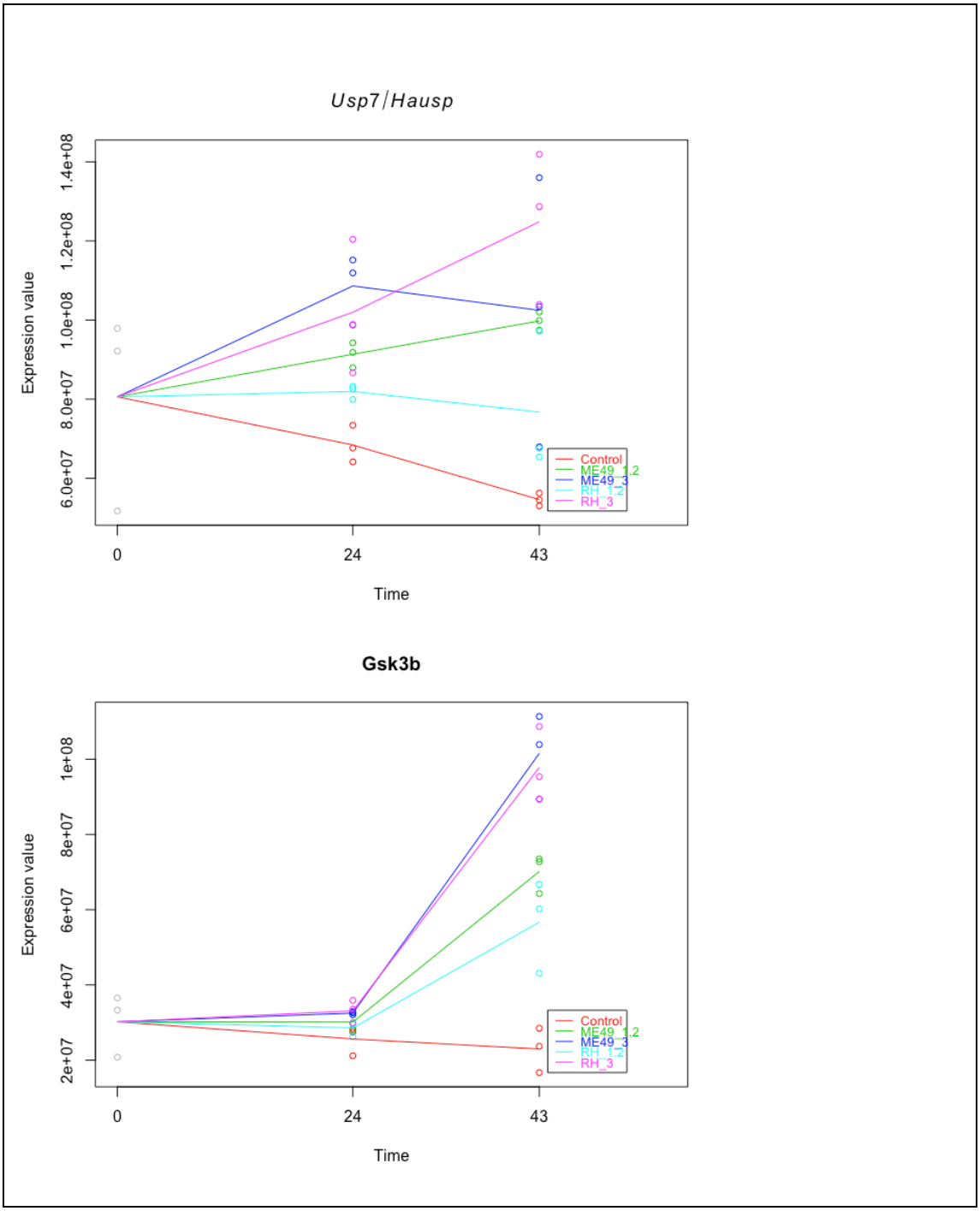


Figure 6.29, continued. Expression profiles of genes related to *Trp53*.

NFkB Pathway

Given the somewhat confusing picture that emerges from NFkB in *T. gondii*-infected cells (see **Chapter 1**) and reports that, in Type II strains at least (17) upregulation of NFkB is wholly dependent on a secreted parasite factor (GRA15), I sought to look at the transcriptional profiles of all NFkB subunit family-members. Moreover, members of this family appeared numerous times in the functional analysis performed in **6.3**.

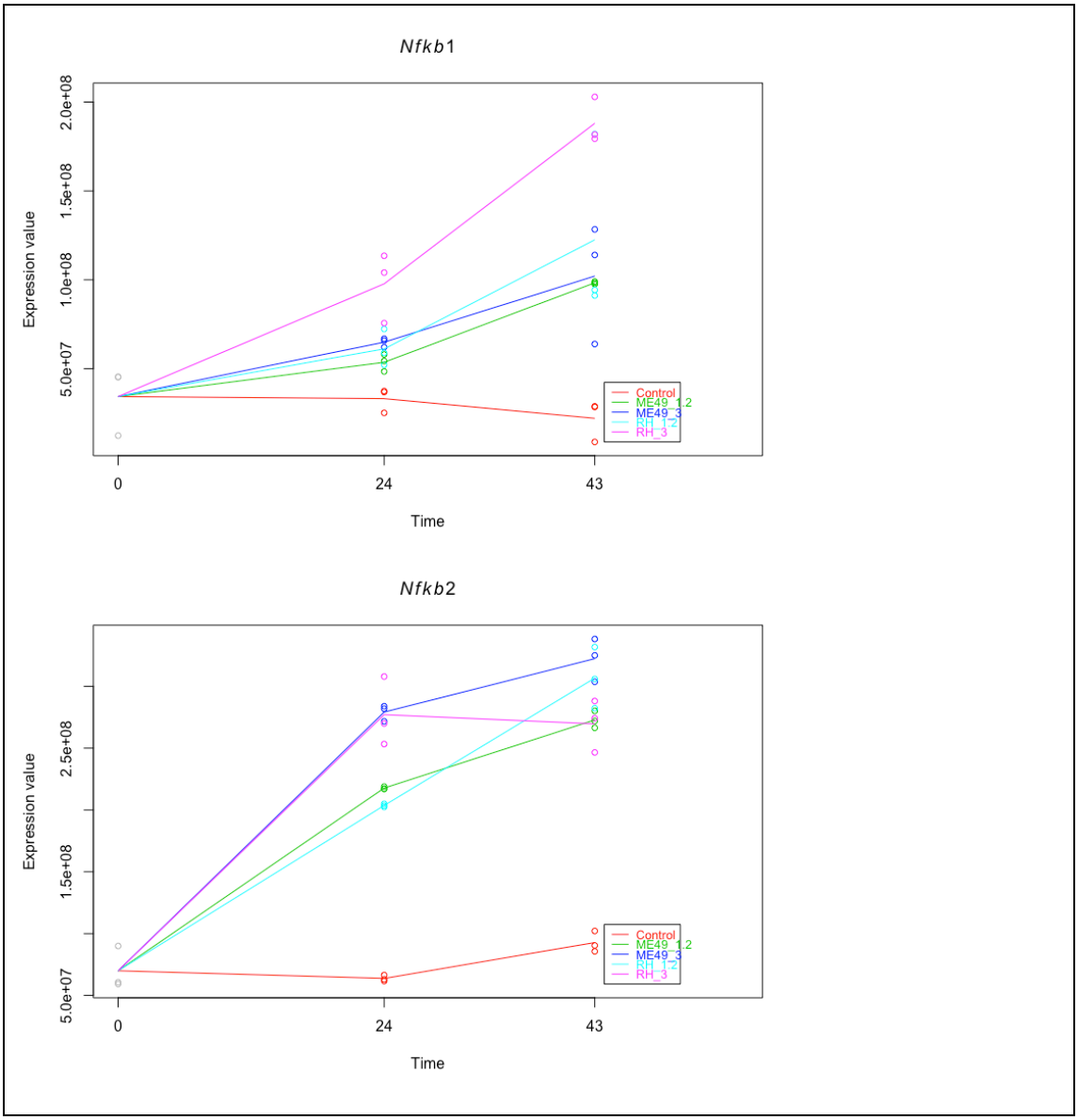


Figure 6.30. Expression profiles of NFkB family members

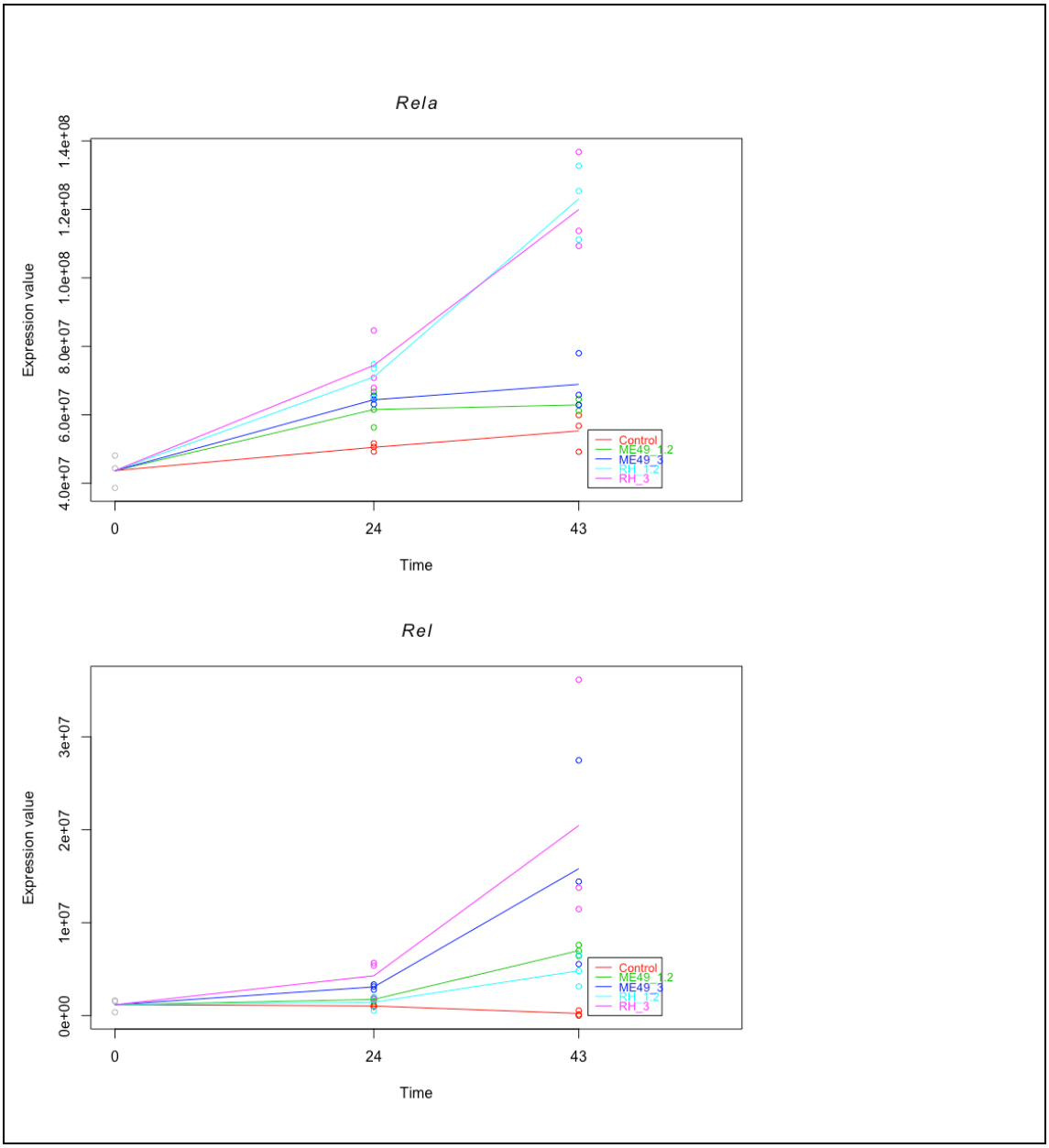


Figure 6.30, continued. Expression profiles of NFkB family members

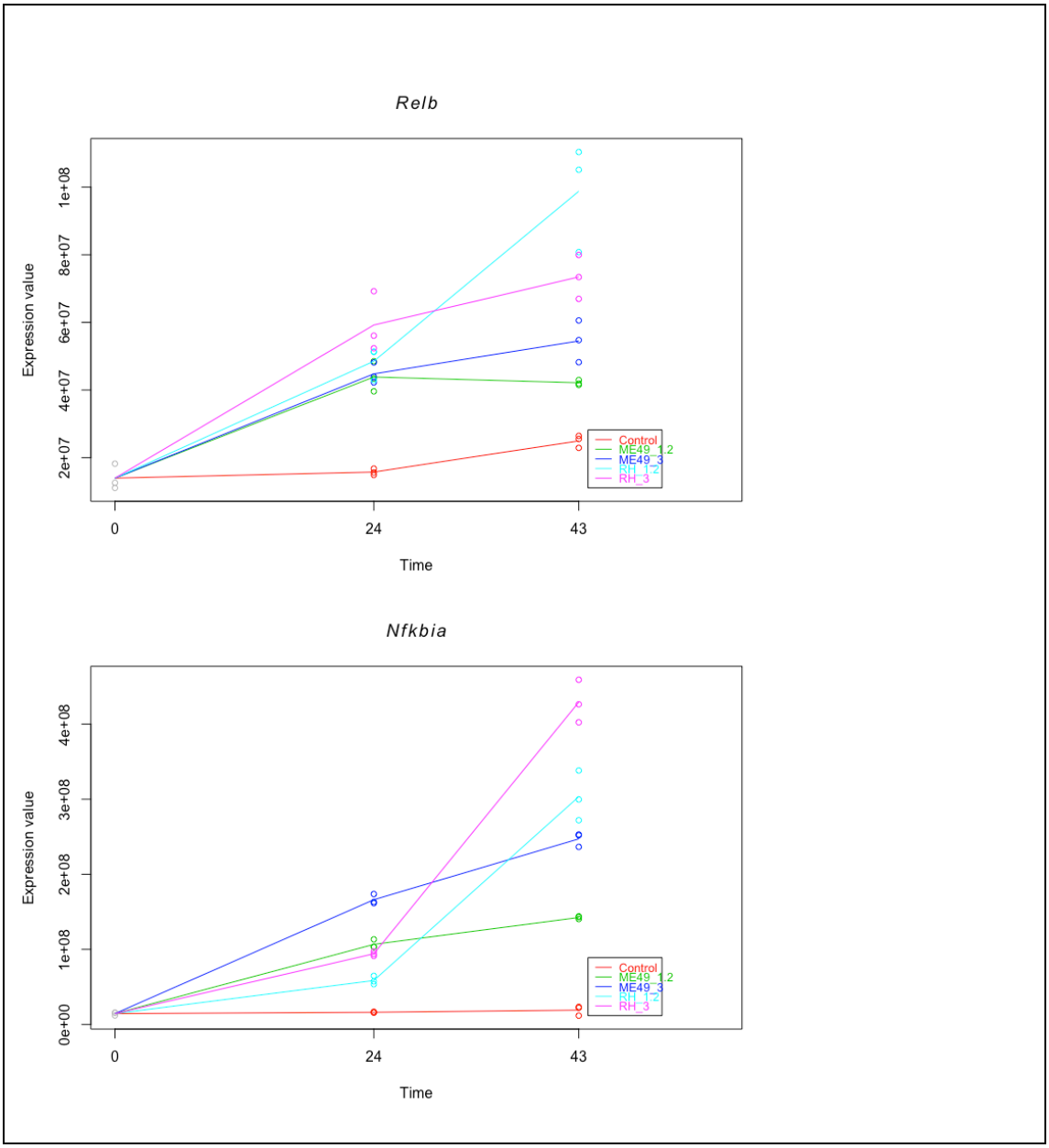


Figure 6.30, continued. Expression profiles of NFkB family members

"Reverse Warburg"

While researching concepts surrounding metabolism in cancer (see **6.4, Discussion**) I came across the intriguing concept of the "Reverse Warburg" effect, the primary mediator of which is Caveolin 1 (*Cav1*).

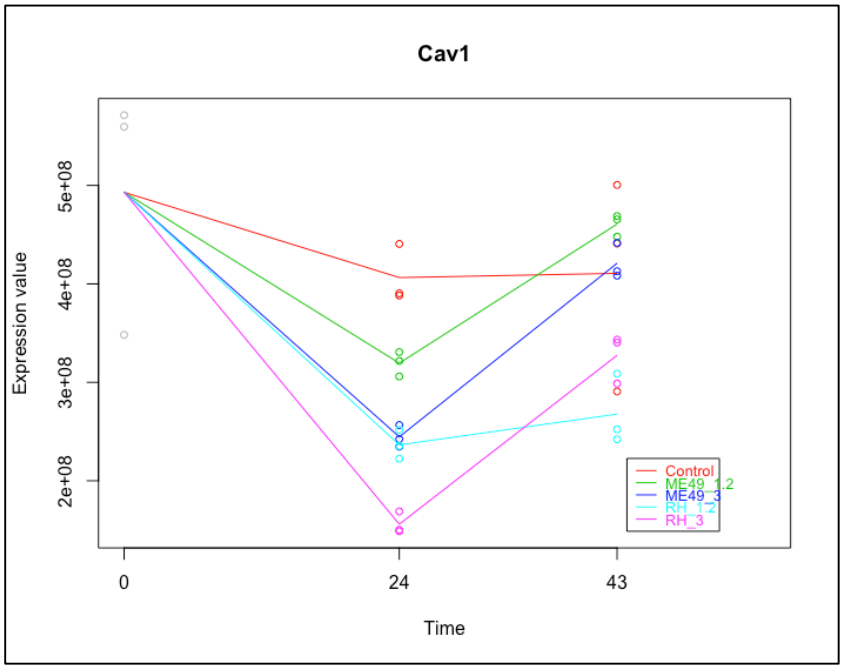


Figure 6.31. Expression profiles of caveolin 1

6.3.4 Lactate Assays

Seeing as many of the genes and pathways that were upregulated related to cancer or glycolysis, I performed a biochemical lactate assay on extracellular medium taken from infected (ME49, MOI 3) or uninfected cells, over a time course of 30h (blanked against unconditioned HG-DMEM). A chart of the results is presented in 6.32

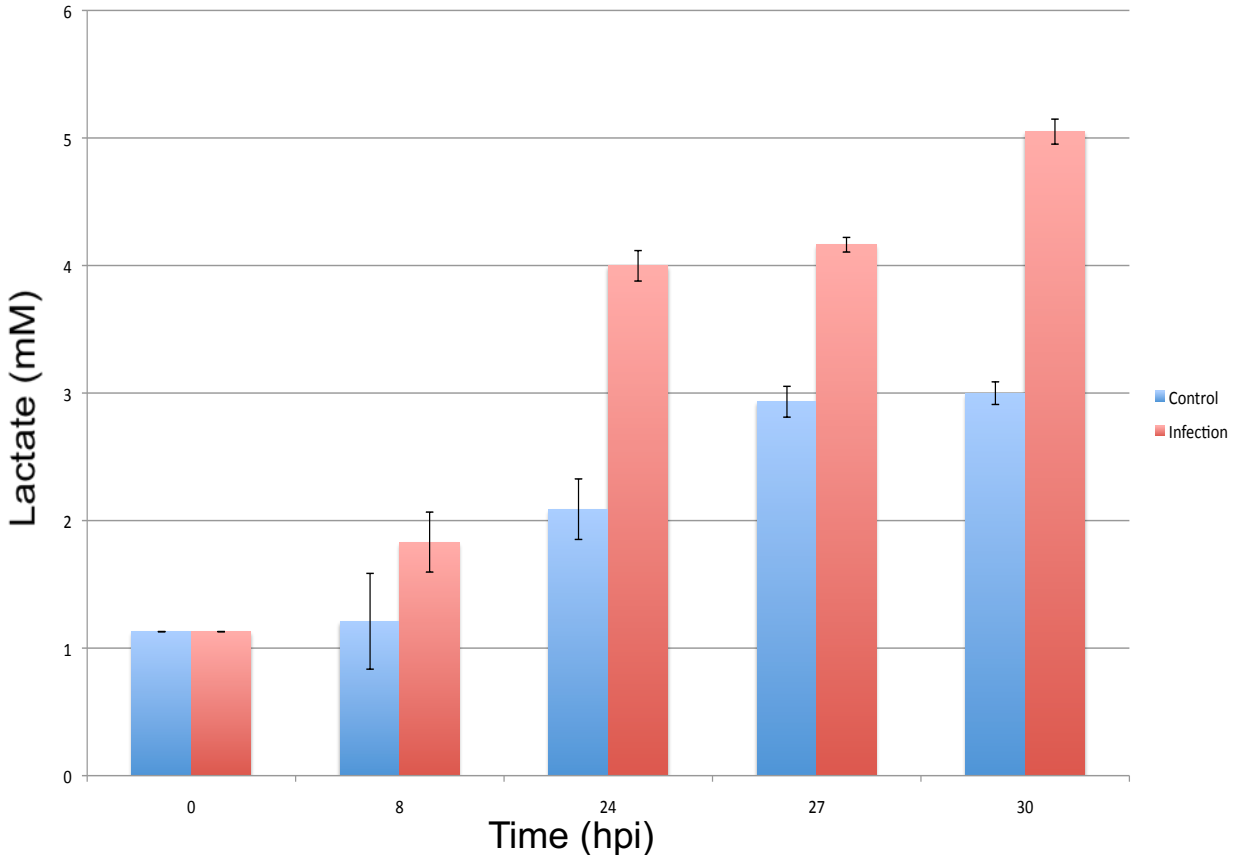


Figure 6.32. Time course of extracellular lactate. Cells were either infected with ME49 at an MOI of 3 (red) or left untreated (blue). Over the course of 30h, the extracellular medium was removed and assayed for lactate.

After a short a time as 24hpi, there is a clear significant difference in the extracellular lactate levels between the infected and control cells, with the former exhibiting a far greater level of lactate.

6.3.5 Western Blots

Given the fact that Hif1a is known to be of particular interest in *T.gondii*-infected cells and is a known mediator of aerobic glycolysis, I infected Hif1a WT and Hif1-/- MEF cells with ME49, at an MOI of 3 for 43 hours, and prepared cell lysates. Then, I extracted lysates from these, with which Bo Shiun Lai performed Western Blots. The aim here was to look at expression of a few key proteins (Figure 6.33). Beta-tubulin was used as a loading control.

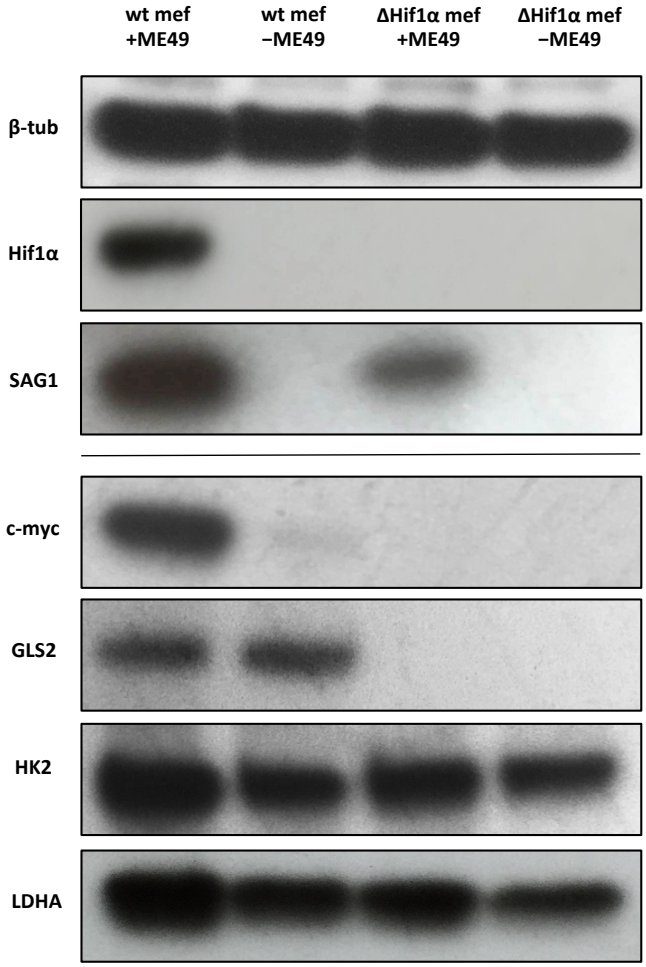


Figure 6.33 HIF1A WT or KO MEFs infected with ME49, MOI3 for 43h.
Loading Control was β-tubulin. Wester blots of the infected samples indicate an abrogation of C-MYC protein expression in the absence of HIF1A, as well as a clear upregulation of HK2 and LDHA, both of which are abated though not totally, in the absence of host cell HIF1A.

As expected, HIF1A levels are completely absent in the KO cell lysates. Similarly, LDHA and HK2 appear higher in the infected cells, though this is

only partially dependent on HIF1A. GLS2 did not appear to vary much between uninfected and infected samples, though its epression was completely abrogated in the absence of HIF1A. If SAG1 expression is taken to be a surrogate for parasite numbers (either through differential replication or invasion), then it is clear that, the parasite contribution is notably lower in the KO cells. While MYC overexpression in infection has been described before (21) and is recapitulated here, perhaps the most surprising result from this series of immunoblots is the total lack of MYC protein in the KO cells.

6.3.6 Methyl Jasmonate

As the first step in glycolysis (aerobic or not), hexokinase 2 is clearly an important enzyme in the host cell metabolic programme initiated by *T. gondii* infection. And, while it is usually cited as being a gene under the transcriptional control of *Hif1a*, it is clear from the western (**6.3.5**) that this is not the entire story. Indeed, the difference in HK2 protein levels between infected and uninfected cells does not appear to depend on the *Hif1a* status of the fibroblast. For that reason, I decided to look more directly at the effect of downregulating the effects of HK2 in a more direct manner, using methyl jasmonate. Methyl jasmonate is a plant stress hormone that has important implications for HK2's role in cancer.

6.3.6.1 Effects of Methyl Jasmonate on Host Cell Numbers

I treated uninfected monolayers grown on coverslips (at 70% confluence) with the same dose of methyl jasmonate as the infected cells, and fixed and mounted them, as in **2.9**. Three fields of view from two coverslips were viewed per condition, and DAPI-stained host cell nuclei were counted. Though there appears to be a slight reduction in cell number in the treated cells, this is still within the SEM.

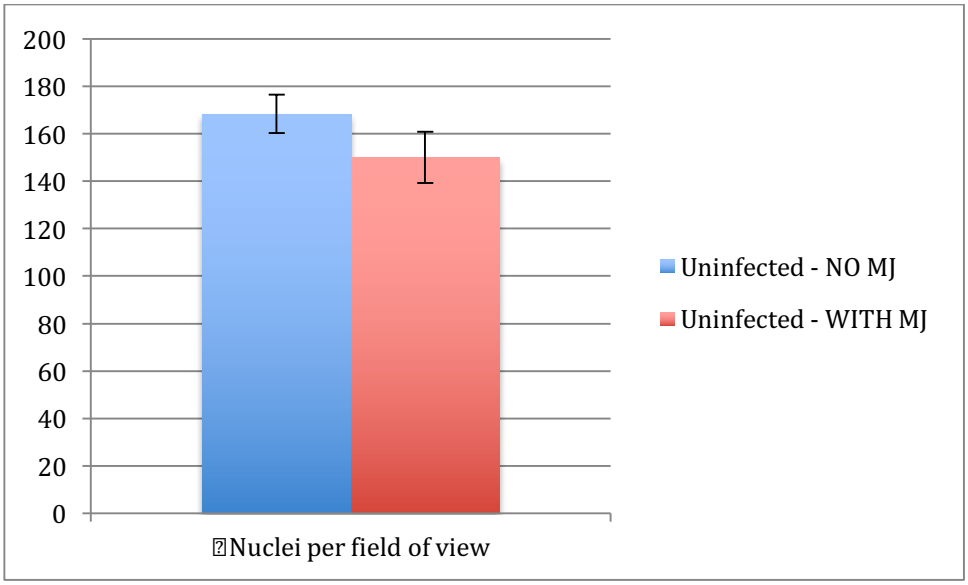


Figure 6.34. Effect of 50 μM methyl jasmonate on host cell numbers.

n = six fields of view from two replicate coverslips/wells. Error bars represent SEM. Cells were seeded with NIH/3T3 cells and treated with methyl jasmonate. No statistically-significant difference emerged between treated and untreated cells.

A representative image that was used for cell counts is shown in Figure 6.27.

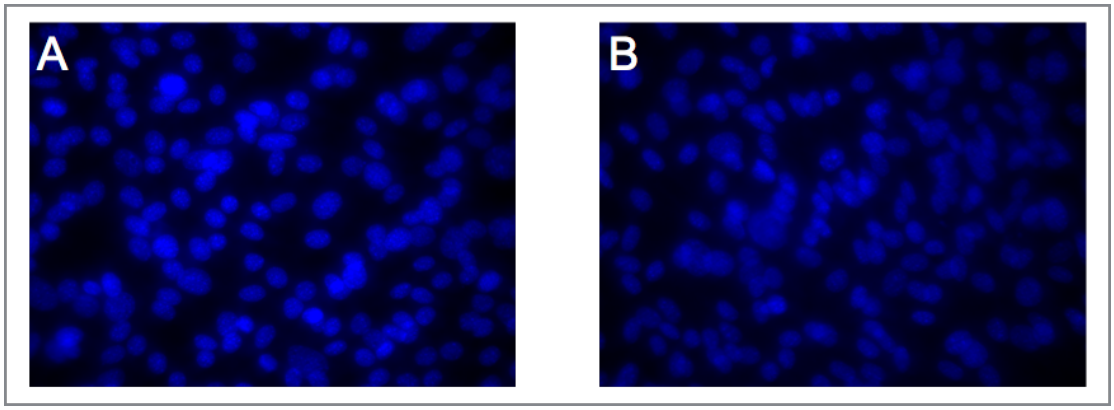
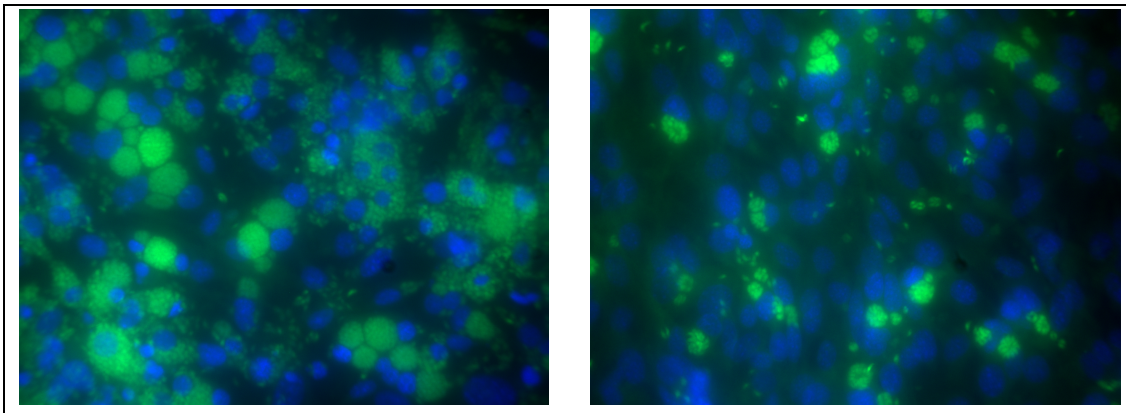


Figure 6.35: Uninfected host cells A) with methyl jasmonate treatment, B) without methyl jasmonate treatment

6.3.6.2 Effects of Methyl Jasmonate on *T. gondii*-Infected Cells

I treated cells infected with either RH, MOI 1.2 or ME49 MOI 1.2 with 50 μM methyl jasmonate and allowed them to grow as usual for 43 hours.

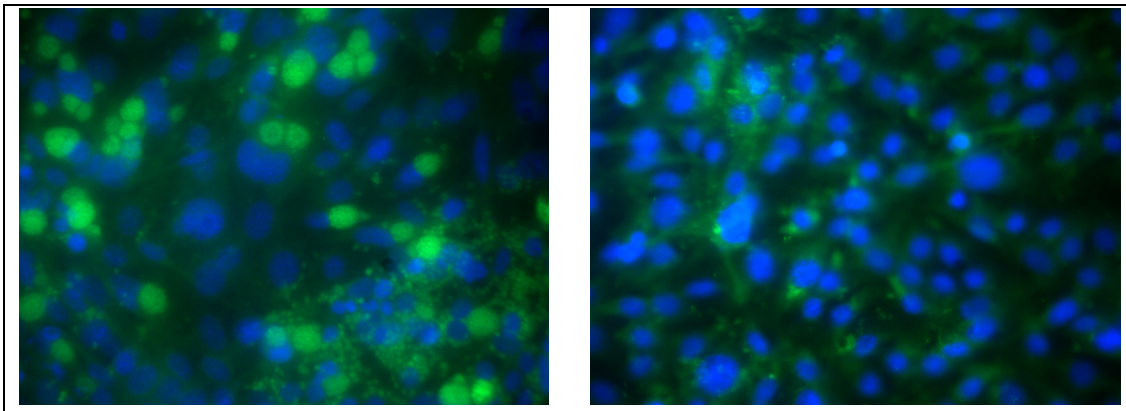
I had hoped to be able to make a quantitative assessment of the effect that methyl jasmonate had on parasite numbers (two coverslips were collected per sample), but the characteristics of the result made this very difficult. Representative images from coverslips of each condition are shown in Figures 6.36 and 6.37.



A B

Figure 6.36. RH-infected cells A) without methyl jasmonate treatment, B) with treatment

In the presence of methyl jasmonate, overall parasite numbers appear lower, and parasitophorous vacuoles appear tighter and less diffuse, perhaps indicating less egression.



A B

Figure 6.37. ME49-infected cells A) without methyl jasmonate treatment, B) with treatment.

In the presence of methyl jasmonate treatment, ME49-containing parasitophorous vacuoles appear smaller, and fewer individual parasites are visible.

When methyl jasmonate treatment was combined with infection with RH, it is clear that a reduction in both PV size and number resulted. However, there was also a large reduction in overall parasite numbers – these may have been from newly-lysed PVs in the untreated sample. This difference in whether parasites appear to be free (in singletons or doublets), within clearly-defined PVs or fully filling the host cytoplasm makes statistical analysis difficult.

In the case of ME49, interpretation is even more difficult. The GFP signal appears extremely diffuse with a few punctae, but no discernible crescent-shaped parasites. Nor where there any visible PVs, of any size. For that reason, I used a higher magnification on the ME49-infected cells to see whether any parasites could credibly be made out but the situation remains very uncertain, Figure 6.38. Nevertheles, given the drastic difference from the untreated sample, it is clear that methyl jasmonate is having an effect – perhaps even a more potent one in ME49 than in RH. A variety of interpretations of these results are discussed in **Chapter 7**.

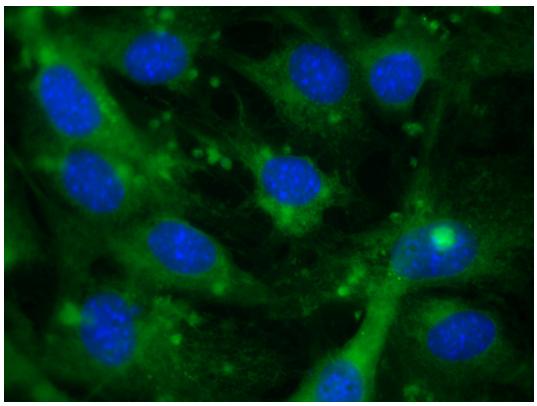


Figure 6.38. ME49-infected cells treated with methyl jasmonate.

6.4 Discussion

It is clear that there a number of different ways to look at transcriptional data, when considering time-courses of several samples. For a start, genes may be dysregulated across each sample's time course, but they may also be dysregulated between the samples. Of course, combinations of these cases are also possible, and care must be taken not to let any one of these factors confound the overall picture. For this reason, I opted to analyse gene expression in my libraries in different methods (over time, between samples) to obtain a broad sense of induced or repressed systems and then looked at them in a more detailed manner by looking at individual gene profiles.

These results that I have shown in **6.3** largely point towards the phenomenon of glycolysis being upregulated as a consequence of infection by *T. gondii*. Notably, this upregulation occurs under normoxic conditions – glycolysis is traditionally thought of as a metabolic phenomenon arising from a restriction in available oxygen. While clearly striking, this is not an unusual phenomenon and, in line with my other observations about numerous cancer-related pathways being upregulated under infection, it is in fact a well-known feature of tumors: the propensity to engage in aerobic glycolysis, even in the presence of plentiful oxygen, the Warburg Effect. The fact that this appears to happen in infection, that is to say in the absence of host cell proliferation and in the absence of host cell genetic transformation, makes it even more striking.

The first inkling that infection by Toxoplasma gondii might have effects on host cell metabolism came with one of the first array analyses of infected host cells (177). These experiments were performed using HFFs infected with a Type II strain (PDS, cloned from ME49). In this study, the authors found that late-stage host cell transcriptional perturbation in infected cells included many genes related to metabolism, notably glycolysis. The authors note a "specific up-regulation of transcripts involved in ***anaerobic***

glycolysis[...]" [emphasis mine], but do not comment on the fact that this apparently 'anaerobic' glycolytic signature of gene expression was being elicited in standard cell culture conditions where oxygen is plentiful. As a result of this apparent contradiction, as well as later studies looking at HIF1A and its interaction with *T gondii*, I examined the phenomenon of ***aerobic*** glycolysis further, through transcriptional profiling of several key genes. In the following discussion, I first look at the genes that emerged from 'parasite-specific' KEGG pathways, and then those that came from uninfected cells (though there is crossover here in terms of cell-cycle regulation). I then turn my attention to the various different facets of aerobic glycolysis that emerge from my data.

6.4.1 Gene and Protein Expression

6.4.1.1 Adhesion and Migration

As well as within parasitology-related pathways, Intercellular adhesion molecule 1 (*Icam1)* and Vascular adhesion cell molecule 1 (*Vcam1*) they were also represented within pathways to do with NK cell mediated cytotoxicity. Interestingly, these genes appear to be highly differentially expressed between the two strains (Figure 6.14), with higher expression levels being achieved in ME49-infected cells. *Icam1* has been looked at largely in the context of the ability of *T. gondii* to cross so-called non-permissive boundaries (such as the blood-brain barrier), given ICAM1's known role in mediating migration of immune cells. Barragan and Sibley note that there exists a strain difference in the ability of Type I and II strains to move across epithelial barriers both *in vitro* and in *ex vivo* matrix simulations, with RH having a far higher migratory capacity (201). This was extended into the observation that Type II strains result in an earlier and greater upregulation of ICAM1 in brain endothelial cells

Vascular adhesion cell molecule 1 has not been studied anywhere near as thoroughly, though it does appear to mediate differences in the inflammatory response observed between C57BL/6 and BALB/c mice, which are described as having different inflammatory cell migration in response to *T. gondii* infection (202). While no causation is shown, this difference is correlated with differential expression of *Vcam1*, with C57BL/6 exhibiting both increased CNS-inflammation, and higher *Vcam1* expression. This study was performed using ME49 only, so strain differences were not possible to compare. Overall, it appears that the differential expression of these two adhesion factors may mediate differences in infection-related immune cell migration, as well as strain-specific inflammatory responses.

6.4.1.2 Cell Cycle

Perhaps unsurprisingly, a number of cell-cyle-related genes were found to be downregulated in the uninfected sample, especially at 43h, at which point contact inhibition will have occurred (Figure 6.18). These are consistent with previous transcriptional studies of contact inhibition (203) and include *Cdc25a*, *Plk1*, *Ccnd1* (also known as *Cyclin D1*) and *Ccna2* (also known as *Cyclin A*). Upregulated genes that have been associated with contact-inhibition include *Cdkn1b* (*p27*) and *Rb1*.

Toxoplasma gondii is known to have an effect on the cell cycle of its host cell (148, 149), though the exact mechanisms of how this is achieved have not yet been elucidated. Both of these studies point towards a parasite-mediated 'push' towards S-phase. This also appears to be the case in my data. For instance, *Ccne2* is strongly upregulated in both strains as compared to the uninfected sample and is required for progression into S-phase. Both studies point to an arrest in cell cycle that prevents infected host cells from proceeding into mitosis, which makes the uptick in S-phase-related processes

intriguing: this is occurring in the absence of cell division and therefore proliferation.

An important gene that I find highly upregulated in all strains at 24hpi (and that is not mentioned either in the *T.* gondii cell cycle studies or in the contact-inhibition study) is *Pcna*, an important mechanistic cofactor for the processivity of DNA synthesis (Figure 6.18) (204). That it is downregulated in uninfected cells is consistent with these being in a state of contact-inhibition while being upregulated in the infected cells, which exhibit an increase in S-phase-related gene expression. Indeed, ectopic overexpression of *Pcna* in contact-inhibited NIH/3T3 cells has been shown to result in unregulated proliferation (205). While obviously infected cells do not engage in unregulated proliferation, it is possible that the parasite-driven 'push' towards S-phase has yet another gene mediating it, that I have identified in the above study.

6.4.1.3 *Toxoplasma gondii* and *Hypoxia inducible factor 1*

Often called the master regulator of the hypoxic response *Hif1* has numerous roles in tumorigenesis, even when oxygen is plentiful. The usually consistutively-expressed alpha subunit (HIF1A) under normoxic conditions has a very short half-life. This is accomplished by the oxygen-dependent action of prolyl hydroxylases (PHD1-3) which hydroxylate the subunit on one or both of it's proline residues. Upon this hydroxylation, HIF1A is targeted for proteosomal degradation, via the E3 ubiquitin ligase (and tumor suppressor) VHL, the von Hippel-Lindau protein. The PHDs that mediate this degradation are themselves sensitive oxygen sensors (absolutely dependent on oxygen as well as iron) and so, in hypoxic conditions, they can no longer hydroxylate HIF1A. Absent this proteosomal targeting, HIF1A is then free to translocate to the nucleus. There, it can dimerise with its stably-expressed partner HIF1B. The heterodimer, along with cofactors from the p300/CBP

family, then mediates the transcription of a number of hypoxia response element (HRE)-containing genes.

Another hydroxylase also controls the activity of HIF1A, the asparaginyl hydroxylase FIH1 (*Factor Inhibiting Hif1*), which was indentified by a yeast two-hybrid screen in 2001 (206). It too is oxygen-regulated, but instead of targeting HIF1A for proteosomal degradation, hydroxylation by FIH1 instead inhibits association of the subunits with their p300/CBP cofactors. That each of these hydroxylases depends absolutely on molecular oxygen, renders them (and therefore HIF1A) sensitive monitor of cellular oxygen levels. However, they also respond to other cues, which may be important indicators of the cell's metabolic state including, importantly, α-ketoglurarate, the first product of reductive glutamine metabolism, often then used for lipogenesis (207, 208). These other interactions (including those with *Myc*) are discussed later on.

The first few transcriptional targets identified as being under HIF1A control were glycolytic enzymes, such as PGK1 and PKM2 (209). Since then, both functional studies involving HIF1A null cell lines, as well as computational studies using the identified "NCGTG" HREs have yielded numerous other genes thought to be at least partially under *Hif1a* control (210, 211).

While having identified glycolytic processes as being upregulated following *T. gondii* infection in their array experiment, Spear et al did not concentrate on the *aerobic* conditions under which this appeared to occur, focussing instead of the specific role of *Hif1a* under "physiological conditions" (i.e. hypoxia[16]). As such, they proceeded to investigate this transcription factor and its possible role in infection further (144). Here, they discovered that the parasite was able to modulate the stability of HIF1A: protein levels were increased upon infection, and appeared to do so in a time-dependent

[16] Though, apart from within large tumor masses it is often very difficult to assess what "physiological" actually represents in terms of oxygen tension.

manner. A later study from the same lab (138) showed, via northern blot, that *Hif1a* RNA was also increased upon RH infection. While the authors' conclusions were focused almost entirely on these effects in hypoxic conditions (less than 3% O_2), the data themselves show a significant HIF1A stabilisation at normoxia (21% O_2) as well. Most significantly, the replication defect that was observed in HIF1A-KO cells under hypoxic conditions could also be seen (although more modestly) at 21% O_2, though this was not commented upon in the publication. This supports the transcriptional data that I have analysed, which show that the phenomenon of aerobic glycolysis is a strain- and MOI-independent phenomenon – suggesting therefore that it is a more generalised feature of tachyzoite infection.

Taken with the Spear microarray data (144), it is then likely that the aerobic glycolysis phenomenon is at least in part responsible for permitting *T. gondii* replication within host cells. Similarly, if SAG1 expression in my western blots (Figure 6.33) are taken as indicative of parasite replication in HIF1A-KO cells, it is clear from my results that the parasite burden in cells lacking HIF1A is much lower – *even under conditions of normoxia.*

From my study, the HIF1a-regulated genes of particular interest are *Hk2*, and *Ldha*, key modulators of glycolysis (though both have also been implicated as targets of MYC possibly in co-operation with HIF1A (212, 213)). Both are highly upregulated upon infection with either strain of *T. gondii*, regardless of MOI, and this is reflected at the protein level too (Figures 6.20 and 6.33). Notably, the B-isoform of lactate dehydrogenase (*Ldhb*) was found to be highly downregulated in infection – a result that is entirely consistent with its role in catalysing the reverse reaction of lactate back to pyruvate (Figure 6.19). Interestingly, in the case of infection, these two genes appears not to be wholly under the control of HIF1a: even in MEFs knocked-out for that transcription factor, an upregulation at the protein level (though abated) is still apparent (Figure 6.33). This suggests other

mechanisms for the control of one or more of the battery of genes I have identified as being important for aerobic glycolysis during *T. gondii* infection. Along with *Hk2 and Ldha*, several transporters crucial to glycolysis are also upregulated in infection (Figure 6.21): *Slc2a1* (also known as *Glut1*), which facilitates entry of glucose into the cell, and *Slc16a3* (also known as *Mct4*) which mediates export of lactate that is produced as a result of glycolysis. Both are frequently upregulated in several cancers and they are considered to be potential drug targets (193, 214, 215). While the regulation of *Mct4* appears to be wholly dependent on HIF1A (216), this is not necessarily the case for *Glut1*. While a role for HIF1A-dependent transcription has been implicated even under normoxia (217), it is unlikely to be the only regulator. Instead, the expression of *Glut1* has also been shown to be a directly target of MYC (218) and also the TRP53 (commonly known as P53) protein, the latter of which apparently mediates its transcriptional repression (219). Notably in my dataset, *Trp53* is itself highly downregulated – perhaps echoing the fact that this tumor suppressor is frequently mutated in cancer – and *Myc* is highly transcriptionally upregulated in the infected samples (Figure 6.27). Interestingly, while the lactate exporter *Slc16a3* is under direct HIF1A control, its family member *Slc16a1* (also known as *Mct1*, another potent lactate exporter) does not appear to be, given that it lacks an HRE in its promoter (216). Instead, *Slc16a3* has been shown to be a direct target of MYC (220), indicating yet another point where MYC and HIF1A may be acting in concert to promote a programme of metabolic reprogramming towards glycolysis.

6.4.1.4 Hexokinase 2 and Sirtuins

For *Hk2*, a great deal can be learned from the cancer literature, especially the work of the Pastorino, Guha and Pedersen groups. This enzyme catalyses the first step of glycolysis, the phosphorylation of glucose upon entry into the cell.

At least four isoforms of this enzymes have been identified and characterised: HK1, HK2, HK3 and HK4. While all are capable of phosophorylating glucose and are inhibited by the product of this glycolysis (glucose-6-phosphate), they vary in terms of their affinities, cell type expression and localisation, and it has been postulated that the localisation and activity of these isozymes have a large part to play in regulating cellular glucose metabolism. HK1, HK2 and HK3 have been reported as having a far higher affinity for glucose (221) with HK4 (also termed Glukokinase) being largely expressed in non-tumorigenic liver and pancreas but upon tumorigenisis, being silenced, yielding instead to higher expression of the other three hexokinases. In my dataset, only *Hk1* and *Hk2* were found to be significantly differentially expressed. Hexokinase 2's prevalence in numerous tumors and cancer cell types (222–226), as well as its specific importance to aerobic glycolysis have led to it being termed the "cancer hexokinase", while *Hk1*'s role has most often been described in terms of its interactions with Akt.

One of the first experiments identifying hexokinase as important for aerobic glycolysis in cancer metabolism was conducted in rat hepatoma cells (222), where – in a stunningly elegant experiment - the authors found that, upon growth in galactose medium (rather than the usual glucose), the formation of lactate was highly reduced, despite those cells' ability to grow well in medium supplemented by either hexose. The authors concluded that this was due to the ability of galactose to bypass the hexokinase step altogether, and notably, that this hexokinase was enriched in mitochondrial fractions. Importantly this and subsequent studies indicated a lack of glucose-6-phosphate inhibition, but only when HK2 is bound to the outer mitochondrial membrane – this is thanks to its N-terminal hydrophobic domain. In ascites carcinoma cells, the localisation of HK2 was also found to preferentially access mitochondrial ATP (rather than having diffused to the cytosol) also contributing to the rapid breakdown of glucose to lactate.

Hexokinase 2 is localised to the mitochondrial voltage-dependent anion channel 1 (VDAC1) whose expression itself is increased at the transcriptional level in all infection states that I assay above. The interaction between HK2 and mitochondria is also thought to be mediated by Peptidylprolyl isomerase D (more commonly known as Cyclophilin D), which is also overexpressed in HEK293 and glioma C6 cells. The role of cyclophilin here may be related to apopotosis as well as glycolysis, as it has been shown that siRNA knock-down of Cycliphilin D releases HK2 from mitochondria, with concurrent increase in apoptotic markers. So, as well as access to ATP and increased glycolytic pathways, it may be that HK2's strategic positioning in cancer cells may function as an anti-apoptotic actor too – some postulate by physically blocking access of actors such as BAX to the mitochondria (227). Confusingly, nuclear localisation of HK2 has also been described in HeLa cells, a localisation which appears to increase upon blockage of enzymatic activity (by, for instance, treating cells with the non-metabolizable glucose analogue 2-deoxyglucose). The purpose of this localisation remains as yet unknown, as do the mechanisms of translocation (though nuclear exportins have been implicated) (228).

Whether this is the case in *T. gondii*-infected cells has been explored by Menendez et al, but only in the context of the Type I strain, RH (229). Here, it appeared that – unlike in cancer – infection did not correlate with mitochondrial localisation of HK2, which may suggest a separate mechanism for hexokinase 2 action here. However, that work was conducted solely in Type I RH. Intriguingly, Type II strains are known to themselves be mitochondrially-associated so it will be critical to examine the mitrochondrial localisation of *Hk2* in Type II-infected host cells. Moreover, nuclear localisation has not been addressed at all.

Another important actor in the role *Hk2* also emerges from the cancer literature: sirtuin 3. Sirtuin 3 is one of seven sirtuins, which are class III

NAD+ dependent histone deacetylases. Sirtuin 3 is localized to the mitochondrial matrix, and activates acetyl-CoA synthetase 2 and glutamate dehydrogenase (GDH), an example of its role as an enhancer of TCA. In a echo to Bustamante et al's early experiments, Shulga et al (230) examined the growth of a variety of cells in galactose (rather than glucose). This resulted in a gradual change of *Hk2* localisation (as assessed through western blots of different cellular fractions) from mitochondria to the cytosol, as well as a concurrent increase in the expression of mitochondrially-associated SIRT3. Through siRNA experiments, the authors found that this effect was mediated by *Sirt3*, via the deacetylation of C*yclophilin D*. Additionally, when *Sirt 3* is knocked-down by siRNA, the usual reduction in membrane potential that follows the transfer of cells to galactose-medium is also abrogated.

Indeed in MEFs, a loss of SIRT3 (via a null mutation) has been found to exhibit an increased glycolytic profile, as assessed by a battery of tests including cell proliferation, lactate production and LC-MS (231). The authors then explored the interplay between *Sirt 3* and *Hif1a*, by looking at HIF1A protein expression in nuclear extracts of SIRT 3 null cells. Increased expression in this case was obvious, and even more pronounced under hypoxia. Conversely, upon SIRT3 overexpresson, *Hif1a* was not as overexpressed, under hypoxic conditions. Further experiments point to the PHDs as the locus of SIRT3 regulation of HIF1A.

On the other hand, *Sirt3* is also cited as a mediator of apoptosis (232), by blocking BAX translocation to the mitochondria (albeit in a mechanism different to that proposed for HK2), in non-transformed cardiomyocytes (233). So, it will be curious to understand the localisation and actual end effect of SIRT3 in *T. gondii*-infected cells between its pro-oxidative phosphorylation, anti-glycolytic, ROS-protective and anti-apoptotic functions.

In my data, while *Sirt3* is ultimately underexpressed in both strains compared to the uninfected host, there appears to be a difference between the

strains and MOIs, with ME49-infected cells exhibiting a definite downregulation at 24hpi (especially MOI 3) whereas RH seems to 'recover' transcription (Figure 6.22). The role of *Sirt3* in cancer is still quite controversial, and thus may undertake different functions under infection by the different strains.

More clear is the pattern found looking at *Sirt6* transcription in infection (Figure 6.22), where an upregulation is seen in uninfected cells, but in none of the infection scenarios. Unlike *Sirt3*, *Sirt6* is nuclearly-localised, and acts as a histone deacetylase on H3K9. This has resulted in reports of NF-kB target silencing. Zhong et al (234) found that, in SIRT6 KO MEF cells, glucose uptake was heightened, via the overexpression of the glucose transporter *Glut1*. Importantly, the authors found co-precipitation of SIRT6 and HIF1a, indicating a potential co-repression by that particular sirtuin (this interaction was not found with, for instance, SIRT1). Given the incomplete picture of *Hif1a* regulation in *T. gondii*-infected cells, it may be that the sirtuins (3 or 6, or other less-well-characterised ones) may have a part to play. Interestingly, *Sirt2*, whose primary roles are to do with cell-cycle control, not the Warburg effect, followed a similar pattern to the uninfected cells for both strains.

6.4.1.5 The role of Glutaminolysis, Fatty Acid Synthesis, Pentose Phosphate Pathway

As well as increased glucose uptake, another facet of the Warburg effect is a concurrent uptake in the non-essential amino acid glutamine. In my dataset, this is supported by the increased expression of, for instance, the glutamine transporters *Slc1a5* (also known as *Asct2*), *Slc7a5* (also known as *Lat1*) as well as both glutaminase genes, *Gls* and *Gls2* (Figure 6.23). Both metabolise glutamine (routinely supplemented in cell culture medium and also reportedly the most abundant extracellular amino acid *in vivo* (235)) into glutamate and

ammonia. Their tissue localisation have been the source of some debate (for instance the GLS2-encoded "liver type" isoforms were first thought to be restricted to hepatic tissues until they were identified in human pancreas and brain tissue(236)) and so obviously, their expression profiles must be considered with care in a cell culture system. That being said, there is evidence of at-times seemingly antagonistic roles in cancer, with the MYC-regulated GLS being associated with increased proliferation and glutaminolysis in the service of macromolecular synthesis in gliomal tumors, and GLS2 isoforms – under TRP53 control and co-operation – acting as tumour suppressors to suppress ROS (237), a feature that has also been noted in human lung and hepatocarcinomas (238, 239). That being said, these roles of glutaminase in infected cells – where the host cell is cell cycle arrested but where nevertheless the "machinery" of proliferation (macromolecular synthesis etc) must be carefully unpicked. Moreover, it is imperative that the genes' transcriptional signatures be related to both protein expression **and enzymatic activity** – as well as the timings of these – in order to properly assess glutamine uptake and deamidation, given that in my hands at least, the transcriptional increase in *Gls2* was not followed by a substantial change in protein level (Figure 6.33). In **Chapter 6**, I profiled two microRNAs that have been shown to both be under MYC suppression and that themselves suppress GLS and GLS2. I would therefore have expected those miRNAs to have been underexpressed but in my dataset, this was not the case. Indeed, while miR-23a's profile was not striking, miR-23b seemed to be upregulated. This highlights that there may be many mechanisms by which GLS and GLS2 are regulated, and that there may be crosstalk between several pathways.

Interestingly, as shown in the western blot, GLS2 appears to be expressed even in uninfected WT cells – that is, even in the absence of MYC expression (Figure 6.33). This further supports the idea that, under conditions of infection, it is likely that *Myc* and (miR-23a/b) are not the only – or even

the major – means of regulating *Gls2*. On the other hand, removal of *Hif1a* does have a large effect on GLS2. This is particularly curious, since in at least two previous studies which sought to computationally identify hypoxia-related target genes (based on analysis of conserved binding sites and prediction of promoter motifs) did not identify *Gls2* as a potential HIF1A-responsive target (210, 211). That being said, the latter of these studies did identify *Gls* but noted that it did not have a consistent response to hypoxia across the datasets the authors were considering and so did not take it further for study).

In tumor cells it is clear that there is a large need for anapleurotic reactions – be these for nucleic acid synthesis, or fatty acid or lipid synthesis. It is this requirement that underpins the idea that glutaminolysis and the production of NADPH as a reducing agent, are key drivers of anapleurosis, with increased glycolysis acting as a source of intermediates, an effect often referred to as "glutamine addiction". The routes for glutamine following import into the cell are several: towards nucleotide synthesis, that of non-essential amino acids (alanine and aspartate), the provision of reducing equivalents and, importantly, the glutathione maintenance.

DeBerardinis et al (240) found, through ^{13}C-NMR labelling experiments that, in the glioma cells they were looking at, glutamine was taken up and shuttled through towards biosynthetic pathways (such as towards aspartate, and oxaloacetate as an anapleurotic carbon within TCA), yes, but that this was greatly in excess of being able to provide the building blocks for such processes themselves. Rather, excessive glutaminolysis, the authors suggested, could be in the service of generating large amounts of NADPH, the electron donor for fatty acid synthesis, but also other anapleurotic processes.

This model suggests a co-operative mechanism by which glucose and glutamine both contribute to the anapleurotic processes of highly proliferating cells. Of course, parasite-infected cells are well-known to be cell cycle arrested (148, 149), and so the macromolecular requirements of the host cell will

clearly be quite different. That being said, despite host cell quiescence, there is of course a highly-proliferative cell within – the parasite itself, within its parasitophorous vacuole. Thus, it may be that *T. gondii* acts as a sort of intracellular tumor, pushing its host cell towards increased anaplerosis, as a means of itself scavenging the necessary intermediates for replication and growth. The role of glutamine may play an important part here, as it appears to do in cancer. For instance, as calculated by Vander Heiden et al (241), the metabolic requirements for the synthesis of fatty acids, amino acids and nucleotides are far greater in terms of reducing equivalents than just ATP: there is "35-fold" asymmetry in the NADPH:ATP ratio. As such, the "efficiency" of oxidative phosphorylation does not come to bear in this unique anabolic scenario, and this is largely dependent on glutamine.

While the two glutaminases and the glutamine importer *Slc1a5* were clearly transcriptionally upregulated, another transporter – *Slc7a11* – points to yet another important role of glutamine: control of ROS (Figure 6.24). Glutamine is one of the key components of glutathione, which acts as a potent intracellular buffer against oxidative stress. While glutamine is crucial to this, it is in great excess in the cellular context (even physiologically), whereas cysteine/cysteine is the limiting component of GSH and can only be imported via the SLC7A11 glutamate/cystine antiporter. Thus, it appears that high levels of glutamine can be imported into the cell and then metabolised into glutamate, which itself is then exported, in exchange for cystine. Then, cysteine (from reduced cystine) together with excess intracellular glutamate can form GSH, via the rate-limiting glutamate-cysteine ligase (GCLC) whose transcript, as with the glutamine and glutamate/cystine transporters is upregulated in infection. As well as these cues that seem to point towards glutamine uptake and glutathione synthesis, the converse pathways appear to have been unequivocally downregulated in infection. For instance, the "Glutathione Metabolism" KEGG pathway made an appearance in nearly all

infection scenarios, with the constituent genes being largely glutathione *S*-transferases (GSTs), which act as detoxification enzymes coupling the metabolism of xenobiotics to GSH. Notably, the majority of cancer research in the area of GSTs has to do with increased expression contributing to chemoresistance of tumors (242) but, in the context of *T. gondii* infection, it may be that transcriptional downregulation of these enzymes contributes to protecting the cellular pool of GSH from spurious (or baseline) GST-mediated depletion.

Another controversial actor in this area is *Isocitrate dehydrogenase 1*, which was also very significantly downregulated in all infections, as well as a repeated contributor to the "Glutathione Metabolism" KEGG pathway[17]. The enzyme encoded by *Ihd1* catalyses the conversion of isocitrate to α-ketoglutarate, in an NADP+-dependent fashion. Mutations in *Idh1* are commonly-associated with gliomas, but much of the discussion here has been centered around the neomorphic character of these mutations, enabling the enzyme to reduce α-ketoglutarate to the potential oncometabolite *R*(-)-2-hydroxyglutarate. Indeed, inactivating mutations – arguably the most analogous situation to the downregulation that I observe in *T. gondii* infection (Figure 6.24) – have not been described in cancer (243, 244). That being said, dominant negative *Idh1* mutations have been seen, with the observed effect of stabilising HIF1A (245). This is almost certainly due to the fact that the product of IDH1 catalysis, α-ketoglurarate, is required by the PHDs that, under normoxic conditions, act to promote the degradation of HIF1A. Thus, with lower levels of α-ketoglutarate available, the PHDs are not able to hydroxylate HIF1A. The regulation of *Idh1* expression itself remains underexplored, although it may be that *Idh1* downregulation is a consequence of the other processes that surround tumorigenesis (or indeed parasitism):

[17] The fact that, in 'normal cellular metabolism' IDH1 is normally thought of as contributing to the pool of NADPH that protects again reactive oxygen guards against overinterpretation of KEGG pathways without a further examination of their constitutive genes.

Robbins et al (246) reported that treating mouse epidermal JB6 P+ (tumor promotable) cells with tumorigenic levels of 12-O-tetradecanoylphorbol 13-acetate (TPA) or UVC radiation reduced IDH1 protein levels as well as activity. It is not inconceivable that a similar process may be happening upon *T. gondii* infection. In any case, the downregulation of *Idh1* in infection is consistent with a preservation of cytosolic citrate – which can feed directly into the fatty acid synthesis pathway.

How might the parasite accomplish this broad-based re-modelling of host cell metabolism? Partly through stabilisation of HIF1A as described by others (144), but also via a programme of interference with host cell signalling networks, including protooncogenes and tumor suppressors.

Wise et al suggest MYC as a key effector of glutaminolysis in human gliomas (247). In their study, increased glutaminolysis was dependent upon MYC expression, via an upregulation of LDHA which in turn redirected glucose towards exiting the cell as lactate (rather than as a lipid precursor). As a result, the authors found, instead of 'usual' TCA processes, *Myc*-transfected cells were instead dependent on glutamine as a substrate for mitochondrial intermediates of anapleurosis. Though the authors did not explore a possible *Hif1a* connection, it is possible, as I find below, that this programme of *Myc*-related glutaminolysis is somehow interconnected with a broader hypoxia-like environment resulting from the stabilisation of HIF1A.

6.4.1.6 *Myelocytomatosis oncogene (Myc)*

A striking feature of infection is the high upregulation of *Myc* in response to infection, both at the transcriptional and protein level. As a b-HLH transcription factor, it interacts with its binding partner MAX (whose half-life is thought to be longer, which is consistent with my transcriptional data) to effect a large transcriptional programme with broad implications for development, differentiation, apoptosis and immune function. Most often, it is

highly overactive in transformed cells, with one estimate being that it is overexpressed in 70% of human tumours (248). Many of the pathways upregulated by MYC are to do with anabolic pathways, for instance fatty acid synthesis, nucleotide synthesis with mitochondrially-sourced intermediates. In this sense, MYC operates on a dual armed system: promoting an increase in glycolysis to yields a rapid source of ATP, as well as the ability to synthesise important anabolic intermediates. I will focus here on its interactions with metabolism (via cooperation with *Hif1a*) and its role in glutamine metabolism (looking later at its interaction with *Trp53*).

Overexpression of *Myc* has previously been shown in cells infected with all three strains, in a variety of cell types, at the protein level as early as 1hpi (though their assay of transcription was only performed at 24hpi) (21). Though the timing may differ – the earliest time point I assayed was 24hpi either for transcription or western blot – these results are consistent with what I have found in terms of transcriptional and protein-level induction. The authors of that work also characterise further a potential route for *Myc* induction via MAPK8 (more commonly known as JNK*)* by treating cells with a JNK-inhibitor and finding that this abrogated *Myc* expression, as assayed by western blot. In my results, *Jnk* expression also transcriptionally increases, which may indicate that this is indeed a possible mechanism, but the western blots show another possible route – one that has not previously been described, involving *Hif1a*.

Very surprisingly, my infection of HIF1A-KO cells resulted in total abrogation of MYC protein expression, suggesting a potentially-novel means of regulating this transcription factor, at least in cells infected with *T. gondii* (Figure 6.33). In most studies, *Myc* has always thought of as being an upstream regulator of *Hif1a*, though my result shows that the situation is clearly more complex, perhaps involving many routes of positive and negative feedback. While *Hif1a* and *Myc* have been known to exhibit antagonistic

effects, particularly in the realm of cell cycle control (249, 250), this has not been shown in terms of a reduction in MYC protein levels, rather by antagonistic effects on MYC target genes themselves (via, for instance, displacement on MYC's classic E-Box binding sites due to the sequence overlap between these sites and HIF1A's HREs). In the latter of these two studies, Koshiji et al used a normoxia-stable HIF1a (Adenovirus-ΔODD) and found no difference in the transcription of *Myc*. On the other hand, MYC has been reported to have effects upon HIF1a, most notably post-transcriptionally, by stabilising the protein (251). Clearly this relationship, and its impacts on target genes downstream of one or both of these transcription factors, needs careful disentangling and is more complex than has been previously understood.

Another cellular regulator that has important implications for both cancer and *T. gondii* infection is TRP53, which has been shown to be dysregulated in infection at least in part by the dense granule protein GRA16 (19). The mechanisms of TRP53 are unclear: While infection with wild-type *T. gondii* results in a robust decrease in TRP53 protein levels, infection with parasites where *Gra16* had been ablated resulted in even lower protein levels. Thus it appears that GRA16 acts to stabilise TRP53, indicating a role for other factors in the overall reduction of TRP53 protein seen in infection. Usually, the control of TRP53 levels is examined in the context of ubiquitination and degradation. One pathway for the control of TRP53 is via ubiquitination by Mouse double minute 2 homolog (MDM2) which subsequently leads to degradation. Ubiquitin specific peptidase 7 (USP7, also known as HAUSP) on the other hand acts to stabilise TRP53 by de-ubiquitination it (252), and so it at first appears curious that *Usp7* transcript levels also appear elevated in infection, although fairly modestly (which mimics USP7 protein levels in Bougdour et al's GRA16 data (19)). However, the situation is not as simple as MDM2 and USP7 acting in antagonistic

directions. The deubiquitination activities of USP7 also act upon MDM2 and so a certain baseline level of USP7 is almost paradoxically necessary for the ubiquitin-mediated degradation of TRP53. How this precise balance of USP7 activity is achieved is not yet fully understood. Yet another mechanism of TRP53 regulation is in fact transcriptional, via BCL6 binding specific *Trp53* promoter sites, though the control of BCL6 itself appears to be post-transcriptional (253). A number of studies have shown that the regulation of TRP53 is complex and involves autoregulation by MDM2, where *Mdm2* is itself transcriptionally-activated by TRP53 (254). In my dataset, *Mdm2* is upregulated – despite the downregulation of *Trp53*. It must also be remembered, however, that this important tumor suppressor can also be regulated by miRNAs such as miR-125b, and miR-504 (255).

6.4.1.7 Apoptosis

It is well-known that infection by *T. gondii* acts potently against host cell apoptosis – an obvious advantage for an intracellular parasite. That these same processes are also affected in tumorigenesis should also not come as a surprise, given the uncontrolled and unchecked proliferation of transformed cancer cells. In my dataset, I was able to replicate the transcriptional results found by Kim et al, who noted downregulation of *Bax* (256). However, it is clear that the regulation of apoptosis depends not simply on the transcriptional regulation of the various actors but also post-transcriptional modifications such as phosphorylation and cleavage (for instance of caspases). Thus, it is perhaps more fruitful to look at apoptosis through the lens of the signalling pathways that modulate it.

One mechanism for this is likely to be via NfkB, where it has been documented that protection from apoptosis in infected cells is absolutely dependent on this pathway. The role of NfkB in *T. gondii*-infected cells is itself controversial, where contradictory effects have often been observed,

likely due to the plurality of host cell types used, the timing of assays, the strain of parasite as well as, as noted by Rosowski et al (17), the fact that many such experiments use as their 'baseline' nuclear translocation of RELA (P65) in response to LPS or TNFA, which may mask more subtle activation, as they themselves found in Type I strains. It must also be borne in mind, that Nfkb has a multitude of effects, many of which could be to an intracellular parasite's advantage at different stages of infection. For instance, blocking NfkB to avoid immune detection is potentially desirable in early infection, while inducing NFkB's anti-apoptotic effects would clearly allow for increased parasite survival. While the identification of GRA15 in Type II strain *Toxoplasma gondii* (17) shows that strain-specific activation (or increase in activation) is indeed possible, other studies show activation (albeit modest in comparison to Type II) in Type I strains such as GT1 at 24hpi (257). Moreover, studies involving *Myc* suggest a broader role for NFkB in all three strains (21). In my own dataset, I see a broad upregulation of all NFkB family-members, with the curious phenomenon that RELA and REL appear to be strain-specifically upregulated (Figure 6.30).

6.4.1.8 Reverse Warburg Effect

A different take on the Warburg effect has been proposed by Michael Lisanti and colleagues (258, 259) in epithelial cancers. Their work suggests a role for the stroma in promoting the metabolic changes associated with Warburg. They argue that the changes that accompany increased aerobic glycolysis in epithelial cancer actually take place in the non-transformed stromal cells, as a result of signalling events from the tumour cells themselves. In this model, the stromal fibroblasts are effectively turned into factories for the tumour, exporting lactate and pyruvate to the tumour microenvironment, for subsequent uptake by the tumour cells thus allowing for their own unrestricted proliferation. This hypothesis came from the observation that

caveolin 1 expression was vastly downregulated in several cancer-associated fibroblasts. Proteomic analysis of CAV1-ablated stromal cells and co-culture experiments of breast cancer MCF7 and fibroblasts (260) further revealed the upregulation of eight glycolytic enzymes (including lactate dehydrogenase A and pyruvate kinase M2). Such a model might be applicable for the case of *T. gondii*-infection. After all, even at high MOIs, not every cell in culture will be host to a parasitophorous vacuole, leaving a proportion of uninfected fibroblasts. If these are taken to be equivalent to the "cancer-associated fibroblasts" described by Lisanti, it is possible that the transcriptional changes associated with the apparent Warburg phenomenon in infection actually results from changes in the uninfected population. Indeed, the experiments that I have performed, be they transcriptomic, proteomic or of lactate output always assay the mixed population of infected and uninfected cells. It cannot therefore be ruled out that the increased lactate output may indeed come from uninfected "infection-associated" cells. Indeed, the severe transcriptional downregulation of *Cav1* (Figure 6.31) might indicate that this is indeed the case, given that it is impossible to distinguish whether this transcriptional difference arises from the uninfected or infected cells from the same experimental well. One clue might be the role of MOI here: one might think that, if the 'Reverse Warburg' effect is in operation with transcriptional differences arising from uninfected cells, then those samples with the broadest infection rate (highest MOI) would exhibit the least transcriptional changes, given that the population of *uninfected* cells in those wells would be lesser. While this does not appear to be the case, this hypothesis ignores any 'magnitude effects' that might arise: if indeed *Cav1* downregulation arises in uninfected cells as a result of signalling events in neighbouring infected ones, the magnitude of transcriptional dysregulation might also be affected. Interaction between *T. gondii* and host cell CAV1 was briefly looked at while examining the establishment of the parasite moving junction during

invasion(13). Immunofluorescence experiments in this study showed that, while CAV1 was selectively excluded from the parasitophorous vacuole while remaining visible in the surrounding host cell. However, no adjacent uninfected were shown in those images, thus leaving the question of CAV1 expression there open.

6.4.2 Methyl jasmonate as an anti-parasitic

A chemical effector that is thought to have an effect on HK2 in cancer is the plant hormone, methyl jasmonate. I initially decided to look at this compound due to its reported effects on HK2 (261), my hypothesis being that methyl jasmonate-mediated disassociation of hexokinase from VDAC might mediate a reduction in aerobic glycolysis (and thus parasite replication). However, while the localisation of host HK2 in infected cells is not definitively mitochondrial under normoxic conditions (229) it has also emerged that methyl jasmonate may mediate a whole host of anti-tumour phenomena, including anti-proliferative, pro-apoptotic, ROS-generating effects. It had been known to affect cancer-cell mitochondria since at least 2005 (262, 263), had its mechanism more fully elucidated in 2008 by Goldin et al (261). These researchers found that treatment of CT26 rat carcinoma cells promoted the selective detachment of HK2 from mitochondria. This is thought to have two effects: first, the disruption of aerobic glycolysis in these cells, and secondly, the triggering of apoptosis in those cells (and those cells only), likely to an increase in ROS levels. My experiments (Figures 6.35 – 6.38) present an intriguing first look at the effect that methyl jasmonate may have on *T. gondii* but the very preliminary nature of these results cannot be stressed enough. Far more characterisation is required before beginning to unpick the jasmonate's role here, and this is discussed in **Chapter 7**.

VII. Discussion and Future Directions

7.1 *Toxoplasma gondii* and microRNAs

While microRNAs have been studied for over a decade now, their functions are only now being characterised – a process made more difficult by the multiple roles that they can play, under different cellular contexts. Most often, they have been looked at in the context of development or cancer[18] and a handful of studies have now examined the issue from the perspective of *T. gondii* infection. My results in **Chapter 4** and **Chapter 5** recapitulate many of these (for instance miR-146a) and also give a broader look at the host cell miRNAs modulated by both RH and ME49. While RNASeq technology like Illumina has undoubtedly been a boon to miRNA research, one of its pitfalls is the sheer volume of data that is produced – and that makes functional interpretation rather difficult. To that end, pathway enrichment is often used as a kind of 'filtering' step, helping to narrow down the vast experimental space. This filtering however depends entirely on the quality of the pathway data available, and those have not been optimised to deal with the role of miRNAs as suppressors of genes. So, enrichment might mean repression, and vice versa.

Of the miRNAs identified in my study as dysregulated by infection, a few warrant special mention. The family of miR-199 was found to be highly suppressed in all strains at all times when compared to the uninfected control. That this miRNA family has elsewhere been implicated in the control of hypoxia (172) renders it all the more intriguing, given that the question of how *T. gondii* infection mediates HIF1A stabilisation has not been fully resolved (138). Useful experiments here would be to overexpress the miRNA and perform analyses of HIF1A and its target genes.

[18] Though this may be an artifact of the likelihood that these are the two most-studied aspects of biological science

The interplay between what 'is known' in *T. gondii* infection and non-*T. gondii*-related processes that may be mediated by miRNAs also call for special attention. An example here is the case of miR-23a/b which, given its documented suppression by MYC (164), I would have predicted to also be suppressed in infection. That this wasn't the case raises questions about other means of regulating this miRNA, as well as of regulating MYC.

7.2 *Toxoplasma gondii* infection and the Warburg Effect

As described in **Chapter 6**, many of the pathways and genes identified as being dysregulated during infection had to do with the phenomenon of aerobic glycolysis. While Otto Warburg's seminal paper *On the Origin of Cancer Cells* (264) is the publication most famously associated with the phenomenon of aerobic glycolysis, the years preceding its publication saw several labs (including Warburg's own[19]) already examining the unusual features of cancer metabolism. As early as 1929, for example, Crabtree looked at the energy consumption of eight different type of transplantable mouse tumour and saw an elevation in aerobic glycolysis in all of them, noting "The constant factor is the possession of a high aerobic glycolysis, which, though not specific for tumour tissue, is a source of energy available for uncontrolled proliferation." (265) Warburg however was the one who extended his findings into a hypothesis, postulating that this was due to a defect in these cells' mitochondrial function, but in the intervening years, this has been shown not to be the case. Rather, the increase in aerobic glycolysis is independent of mitochondrial misfunction and indeed, mitochondria do not overwhelmingly exhibit any real defects in cancer cells. Between 1958 and 1964, Wasley D. Yushok published a series of papers (266–268) in which he examined the effects of different metabolisable sugars on the respiration rates of ascites,

[19] For example his 1924 paper *Über den Stoffwechsel der Carzinomzelle* , On the Metabolism of Cancer Cells (283)

with the observation that the ability of hexokinase to differentially phosphorylate them was a key point of potential control.

The difficulty in much of the cancer literature to do with metabolism and the Warburg effect is the inherent genetic heterogeneity that causes tumorigenesis. As such, different tumors will almost certainly have different metabolic needs and so, hypotheses seeking to explain the Warburg effect abound. A few of these include the impact of glutaminolysis, protection from ROS and/or apoptosis, autophagy, mitophagy and, of course, immune-related functions. This leads to an attractive idea: the use of *T. gondii* as a proxy for tumorigenesis that is able to decouple the metabolic effects from the genetic transformation or dysregulated proliferation that is exhibited in cancer.

A preliminary model for the regulation of host cell metabolism by *T. gondii* is shown in Figure 7.1, where the parasite has effects in targeting particular signalling netwkorks and transcription factors.

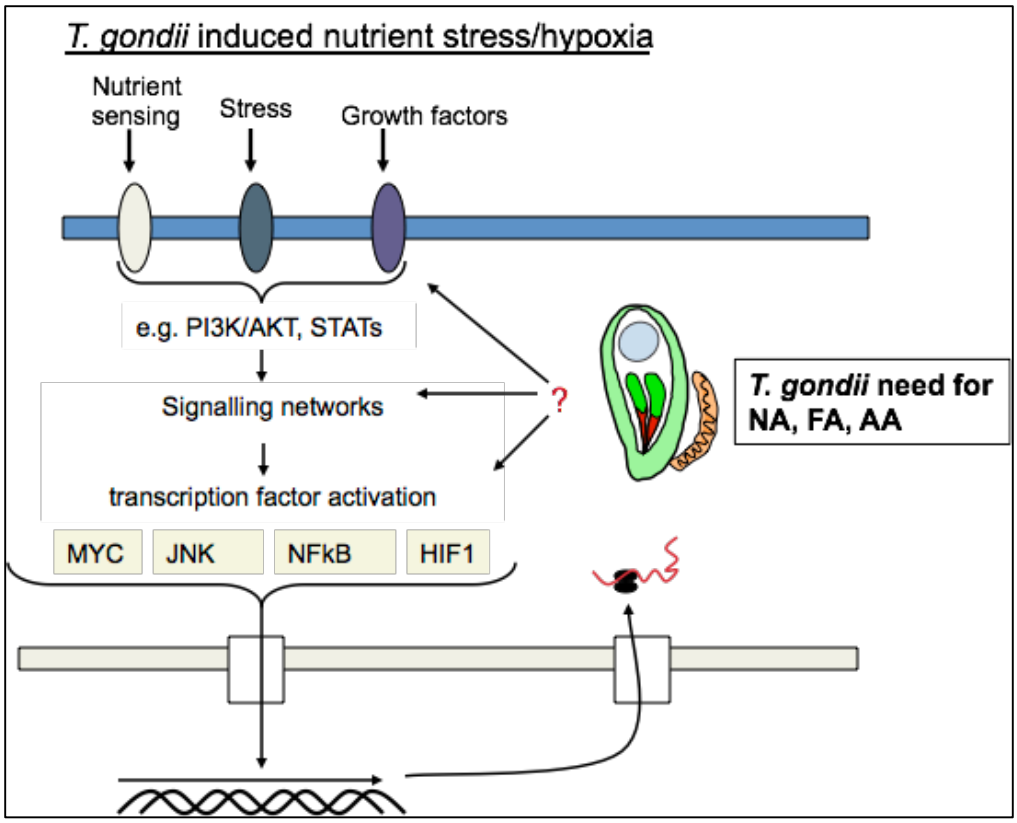


Figure 7.1. A schematic of the key requirements that *T. gondii* has upon intracellular infection of the host cell. Some of the key signalling networks that are known to be modulated by the parasite are noted, as well as potential 'reasons' why it these might promote parasite survival. NA = nucleic acids; FA = fatty acids; AA = amino acids. Adapted, with permission, from Dr J. Ajioka.

This follows from the obvious need that an intracellular parasite has for certain key cellular 'commodities' such as nucleic acids, fatty acids and amino acids. Moreover, there is a great need for reducing equivalents, and so more complex processes such as glutamine metabolism and ROS-protection may also be tweaked. Indeed, it is likely that the parasite is able to modulate the host resulting in changes both at the level of gene expression and metabolite levels. Preliminary studies have looked at metabolite levels in infected cells using ^{1}NMR (Roohi, personal communication) and those results have been used to supplement the gene expression information revealed during my analyses in a more detailed model of intracellular infection, highlighting

differentially-modulated genes and metabolites. Such a model can help form the basis for future experiments on specific pathways.

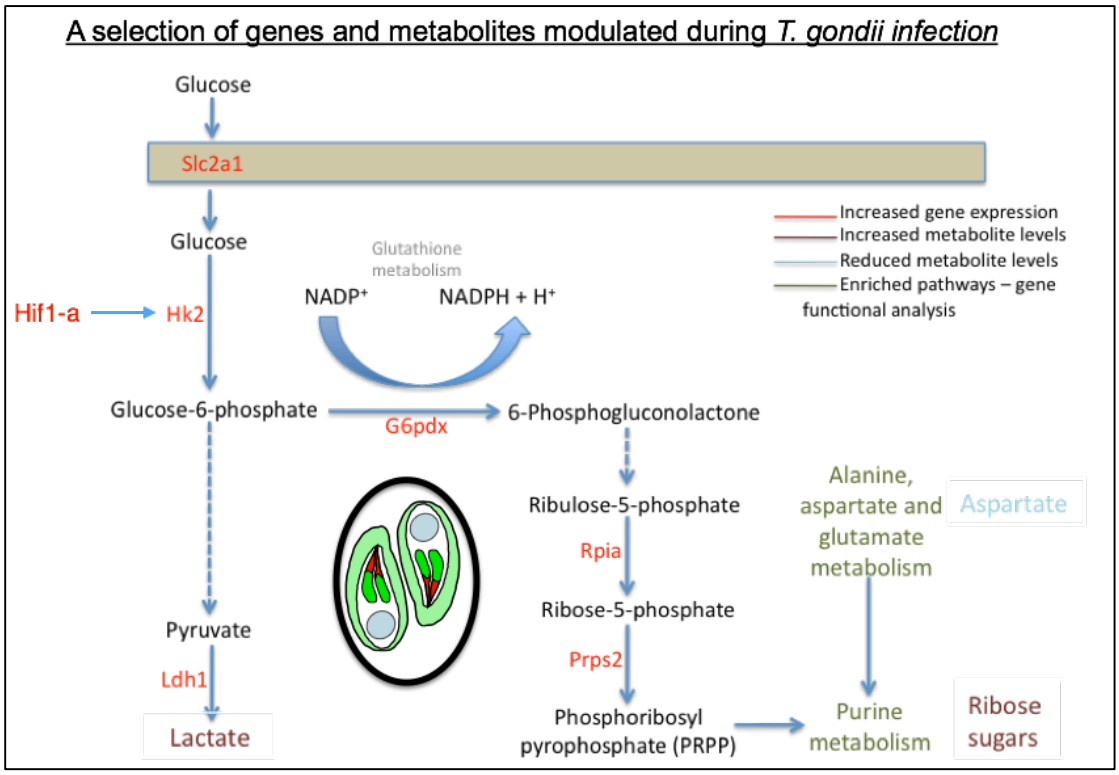


Figure 7.2. A schematic of some of the genes and metabolites modulated during *T. gondii* infection. Gene expression data from Chapter 6 were combined with data communicated to me personally by a colleague regarding the modulation of metabolites. (Adapted, with permission, from Dr. A. Roohi).

That being said, it is increasingly clear that a rather limited number of transcription factors are at the root of this metabolic reprogramming, though these are interconnected in often non-obvious ways. As Yeung et al note "The regulation of energy metabolism can be traced to a "triad" of transcription factors: c-MYC, HIF-1 and p53." (269) – and *T. gondii* infection appears to modulate all three. While MYC (20) and TRP53 appear to have regulatory parasite secreted factors identified the situation is still not clear cut – after all, GRA16 appears to have a stabilising role on TRP53, unlike the wholesale decrease in mRNA that I see, or the protein decrease that Bougdour themselves note (19). It may be that at least some of these effects are

mediated via the NFkB pathway but the parasite effector associated with that regulation is strain-specific (19), and it appears that the overall upregulation in glycolytic/cancer-like processes are not. What is clear is that we need a better understanding of how these signalling pathways intersect under conditions of infection. One of the most striking results from my data suggest that HIF1A may have a regulatory role on MYC (251), whereas the relationship between them has thus far been described as the reverse. As such, a sequential dissection of these three key transcriptional networks, using classical genetics, siRNAs or overexpression studies, will be crucial in understanding what nexuses of control are likely targeted by *T. gondii*.

7.3 Specific Future Directions

7.3.1 Mitochondrial Localisation and Activity of Hexokinase 2

Given the upregulation of hexokinase 2 in infection, taking a page from the cancer literature and exploring its localisation would be a useful path to explore. It has been proposed that, apart from being the 'cancer hexokinase' upregulated as part of the generalised Warburg phenomenon, its localisation to the mitochondria is key to this effect (227, 270, 271). In certain cases, growing cells in medium where galactose is the central carbon source (instead of glucose), has been shown to result in the dissociation of HK2 from mitochondria (222, 272).

The steps to undertake this experiment would be as follows:

1) Ascertaining the growth rate of host cells in the galactose medium, versus the glucose medium to ensure that any growth defect is controlled for.

2) Ascertaining the expression and cellular localisation of HK2 during infection and whether that changes upon growth of infected cells in galactose.

3) Ascertaining the metabolic state of these cells. Ideal for this is a piece of equipment called the Seahorse XF (273), which simultaneously and in real time analyses oxygen consumption (OCR) and extracellular acidification

(ECAR. In a sense, this is a modern interpretation of the biochemical lactate assay that I have performed, but also includes oxygen electrode experiments.

7.3.2 Carbon Labelling Experiment

Apart from specifically assaying particular enzymes such as HK2, a more global picture of host cell metabolism upon *T. gondii* infection could be gained by performing, as Deberardinis et al did with glioma cells (274), 13C-NMR labelling experiments. By measuring which particular carbon is labelled, one can then ascertain the eventual 'destination' of the increased glucose taken up during infection – whether it is to supply the parasite directly, to supplement depleted host nutrients or a combination of both. Not only that, it can – through the proportion of differentially-labelled products - give an idea of the rates of the different routes through central carbon metabolism, and how these are altered upon infection.

7.3.3 Methyl Jasmonate

My results in 6.3.6 provide a very preliminary glimpse into a potential role for methyl jasmonate as an anti-parasitic, and this may give useful insight into the mechanisms by which *T. gondii* reshapes its metabolic host cell niche. Of course, my result in and of itself is inadequate to make any such claims, but there are a number of experiments that could be undertaken to clarify what methyl jasmonate's action is here. For a start, it is crucial to attempt to separate the drug's effects on the host and on the parasite. While it does not seem that methyl jasmonate had a noticeable effect on host cell numbers, live-dead staining (using for instance trypan blue) would confirm that this was indeed the case. Moreover, it will be important to look at the metabolic state of treated cells – this is true of both infected and uninfected cells.

Given that my preliminary results suggest lower numbers of PVs and parasites in treated cells, it is then necessary to examine what the disadvantage actually is: is it a block on invasion? A replicative disadvantage?

This can be sorted out to an extent by modifying the 'order of operations' of such an experiment. These would be the steps to do this:

1. Apply parasites to a host cell monolayer in 'normal' medium
2. Allow the parasites to invade for an hour
3. Rinse the monolayer with PBS to remove parasites that have not invaded
4. Replace the medium with either fresh medium or fresh medium containing methyl jasmonate.

A tight time course following treatment would then follow the timing of replication of the parasites that had been able to invade – distinguishing between PVs containing different numbers of parasites at each time point. Further, if the 'PBS-rinse' were centrifuged and any resulting pellet examined one could also examine the parasites that had explicitly not invaded. Another necessary control is to pre-treat parasites with methyl jasmonate (or fresh medium) for a few hours before using them to infect the monolayer. This is always a little tricky given the obligate intracellular nature of parasites, but it is a reasonable compromise as long as it is well-controlled. Finally, the differentiation potential of treated cells would also need to be controlled as it is possible that by altering host or parasite metabolism, the transition to bradyzoite is promoted. This could be done with simple staining using bradyzoite antigen 1 (BAG1) antibody at different times following.

7.4 *Toxoplasma gondii* and the Warburg Effect: Clinical Implications

7.4.1 Chemotherapy

While methyl jasmonate may (or indeed ultimately may not) exhibit metabolism-based anti-parasitic effects, it is not the only compound to have been considered in both the context of cancer and *T. gondii* infection. In several cancer studies, the compound 3-Bromopyruvate has been employed as an HK2-inhibitor. Interestingly, in one study looking at the effect of this compound in *T. gondii*-infection, the authors found a dose- and time-

dependent reduction in parasite proliferation following treatment of infected LLC-MK2 (rhesus monkey kidney) cells; 3BP treatment had no effect on host cell proliferation (275). While de Lima et al attribute this effect to direct action on parasite metabolism, they do not address at all the possibility that 3BP may instead modulate metabolism of the infected host cell instead, as it appears to do in cancer. It would be interesting to examine this further, through one of the metabolic studies proposed above.

In a clinical context, though, care must be taken using such newly-characterised drugs : 3BP has recently found itself at the centre of a tragic controversy, where its use in an unlicensed alternative therapy clinic is thought to have caused the deaths of three patients (276). The full connection between these deaths and the use of 3BP is still being investigated but the case highlights the need to fully characterise these types of metabolic drugs extremely carefully – their effects are likely to be broader than simply targeting a single metabolic pathway and even if they do, the central position of cellular metabolism (including, for instance the brain's requirement for glycolytic metabolism) makes this a particularly challenging pathway to target therapeutically.

7.4.2 Diagnosis

^{18}F-FDG-PET has, since the 1980s, been a commonly-used technique for the visualisation of tumors *in vivo* (277). It rests on the use of the non-metabolisable glucose analogue 2-fluoro-2-deoxy-D-glucose (F-18), and subsequent visualisation using Positron Emission Tomography – a direct result of the Warburg Effect whereby tumours take up glucose at highly increased rates. This technique has been shown to be useful in distinguishing between lymphoma and toxoplasmosis in patients with AIDS (278). Central Nervous System (CNS) lesions arising from either condition are near-identical in appearance when evaluated through traditional contrast-based imaging

methods such as CT or MRI (279). Drawing upon previous studies (REF) which showed that CNS lymphoma, like most cancers, accumulate ^{18}F-FDG at a high rate, Hoffman et al demonstrated that it was possible to use this hypermetabolic feature of lymphoma in ^{18}F-FDG-PET, to discriminate between CNS lesions arising from malignancy or from infection. Indeed, lesions arising from non-malignant pathologies were shown not to accumulate ^{18}F-FDG as much as those resulting from lymphoma. As a result, ^{18}F-FDG-PET has become, at least in patients with AIDS, a valuable tool to resolve the often confusing differential diagnoses of CNS lesions. However, as Hoffman et al themselves acknowledge, this method is imperfect. In their study, seven of the 11 patients were at the time receiving treatment for toxoplasmosis, though only four of them were eventually diagnosed with it. This is due to the widespread clinical practice of 'empirical treatment' where patients with AIDS who exhibit characteristic neurological symptoms are pre-emptively prescribed an anti-toxoplasma regimen, even in the absence of serological evidence of parasitic infection. Only if symptoms prove refractory might other avenues, such as an FDG-PET scan be pursued. As such, the FDG-PET profile of *T. Gondii* infection in the absence of treatment remains unknown. Moreover, the majority of toxoplasmosis cases in patients with AIDS results from the reactivation of encysted parasites, a situation which may present a different metabolic profile (and thus a different FDG-PET reaction) to acute primary infection.

The clinical FDG-PET response of acute Toxoplasma infection in HIV-negative populations is difficult to assess, since most primary infections are asymptomatic and would not require even a visit to the doctor, much less a PET scan. As a result, any clinical metabolic information in this regard comes from patients for whom an FDG-PET was indicated for other reasons. At least two cases are noteworthy in this respect. In 2001, Sandherr et al reported the case of a patient apparently in remission from Hodgkin's

Lymphoma, whose follow-up full-body FDG-PET scan revealed a splenic locus of high FDG uptake. The presumptive diagnosis was of a recurrence of the lymphoma but, after numerous serological tests, the patient was eventually diagnosed with, and successfully treated for, toxoplasmosis. Importantly, though the authors do not make note of this, the method of diagnosis (positive igm-antibodies rather than igg characteristic of chronic infection) would suggest that this was a primary infection (280).

Spieth et al's patient exhibited enlarged lymph nodes, more than a year after undergoing a wide resection of a scalp melanoma (281). As in the previous case, a full-body FDG-PET was indicated and its results "further strengthen[ed] the presumptive diagnosis of lymph node metastases". However, after histological examinations pointed to toxoplasmosis, acute infection was diagnosed immunologically. In this case, the infection was left to 'run its course' (as is advised in most non-immunocompromised patients) and, six months later, an FDG-PET scan revealed no unusually hypermetabolic loci. Furthermore, serial igm measurements reflected a marked attenuation of acute primary *T. gondii* infection.

In a clinical context, these studies serve primarily to emphasise the need for care when interpreting FDG-PET scans but the bio-energetic picture that they hint at is much less clear. Sandherr et al speculate that the FDG-PET false-positive they observed might be due to "increased glycolysis of activated granulocytes and macrophages" but this remains conjecture and Spieth et al offer no explanation for their result. Either way, the idea that acute primary *T. gondii* infection (as distinct from chronic infection) may, *in vivo*, mimic present with the same, and most characteristic, metabolic feature as cancer is intriguing, and perhaps plant the seeds for parasite-induced Warburg effect *in vivo*.

VIII. References

1. Hill DE, Chirukandoth S, Dubey JP (2005) Biology and epidemiology of Toxoplasma gondii in man and animals. *Anim Health Res Rev* 6(1):41–61.
2. Ajioka JW Dr James Ajioka — Department of Pathology. Available at: http://www.path.cam.ac.uk/directory/james-ajioka.
3. Martin S (2001) Congenital toxoplasmosis. *Neonatal Netw* 20(4):23–30.
4. Petersen E, Dubey JP (2001) Biology of toxoplasmosis. *Toxoplasmosis : A Comprehensive Clinical Guide*, eds Joynson DHM (David HM, Wreghitt TG (Cambridge University Press, Cambridge), p 395.
5. Dubey JP, Lindsay DS, Speer CA (1998) Structures of Toxoplasma gondii tachyzoites, bradyzoites, and sporozoites and biology and development of tissue cysts. *Clin Microbiol Rev* 11(2):267–299.
6. Rougier S, Montoya JG, Peyron F (2017) Lifelong Persistence of Toxoplasma Cysts: A Questionable Dogma? *Trends Parasitol* 33(2):93–101.
7. Lüder CGK, Rahman T (2017) Impact of the host on Toxoplasma stage differentiation. *Microb cell (Graz, Austria)* 4(7):203–211.
8. Sibley LD, Boothroyd JC (1992) Virulent strains of Toxoplasma gondii comprise a single clonal lineage. *Nature* 359(6390):82–85.
9. Khan A, et al. (2011) Genetic analyses of atypical Toxoplasma gondii strains reveal a fourth clonal lineage in North America. *Int J Parasitol* 41(6):645–655.
10. Sibley LD, Mordue DG, Su C, Robben PM, Howe DK (2002) Genetic approaches to studying virulence and pathogenesis in Toxoplasma gondii. *Philos Trans R Soc Lond B Biol Sci* 357(1417):81–8.
11. Saeij JPJ, Boyle JP, Boothroyd JC (2005) Differences among the three major strains of Toxoplasma gondii and their specific interactions with the infected host. *Trends Parasitol* 21(10):476–81.
12. Minot S, et al. (2012) Admixture and recombination among Toxoplasma gondii lineages explain global genome diversity. *Proc Natl Acad Sci* 109(33):13458–13463.
13. Mordue DG, Desai N, Dustin M, Sibley LD (1999) Invasion by Toxoplasma gondii establishes a moving junction that selectively excludes host cell plasma membrane proteins on the basis of their membrane anchoring. *J Exp Med* 190(12):1783–92.
14. Saeij JPJ, et al. (2007) Toxoplasma co-opts host gene expression by injection of a polymorphic kinase homologue. *Nature* 445(7125):324–7.

15. Saeij JPJ, et al. (2006) Polymorphic secreted kinases are key virulence factors in toxoplasmosis. *Science* 314(5806):1780–3.

16. Niedelman W, et al. (2012) The Rhoptry Proteins ROP18 and ROP5 Mediate Toxoplasma gondii Evasion of the Murine, But Not the Human, Interferon-Gamma Response. *PLoS Pathog* 8(6):e1002784.

17. Rosowski EE, et al. (2011) Strain-specific activation of the NF-kappaB pathway by GRA15, a novel Toxoplasma gondii dense granule protein. *J Exp Med* 208(1):195–212.

18. Pernas L, et al. (2014) Toxoplasma Effector MAF1 Mediates Recruitment of Host Mitochondria and Impacts the Host Response. *PLoS Biol* 12(4):e1001845.

19. Bougdour A, et al. (2013) Host Cell Subversion by Toxoplasma GRA16, an Exported Dense Granule Protein that Targets the Host Cell Nucleus and Alters Gene Expression. *Cell Host Microbe* 13(4):489–500.

20. Franco M, et al. (2016) A Novel Secreted Protein, MYR1, Is Central to *Toxoplasma* 's Manipulation of Host Cells. *MBio* 7(1):e02231-15.

21. Franco M, Shastri AJ, Boothroyd JC (2014) Infection by Toxoplasma gondii Specifically Induces Host c-Myc and the Genes This Pivotal Transcription Factor Regulates. *Eukaryot Cell* 13(4):483–493.

22. Parham P (2000) *The Immune System* (Garland Publishing, Current Trends, New York).

23. Aliberti J, et al. (2000) CCR5 provides a signal for microbial induced production of IL-12 by CD8 alpha+ dendritic cells. *Nat Immunol* 1(1):83–87.

24. Pepper M, Hunter CA (2007) Innate Recognition and the Regulation of Protective Immunity to Toxoplasma gondii. *Toxoplasma: Molecular and Cellular Biology*, eds Ajioka JW, Soldati D (Horizon Bioscience, Norfolk), pp 111–126.

25. Suzuki Y, Orellana MA, Schreiber RD, Remington JS (1988) Interferon-gamma: the major mediator of resistance against Toxoplasma gondii. *Science* 240(4851):516–8.

26. Gazzinelli RT, Hieny S, Wynn TA, Wolf S, Sher A (1993) Interleukin 12 is required for the T-lymphocyte-independent induction of interferon gamma by an intracellular parasite and induces resistance in T-cell-deficient hosts. *Proc Natl Acad Sci USA* 90(13):6115–6119.

27. Molestina RE, Sinai AP (2005) Host and parasite-derived IKK activities direct distinct temporal phases of NF-kappaB activation and target gene expression following Toxoplasma gondii infection. *J Cell Sci* 118(Pt 24):5785–96.

28. Shapira S, et al. (2005) Initiation and termination of NF-kappaB signaling by the intracellular protozoan parasite Toxoplasma gondii. *J Cell Sci* 118(Pt 15):3501–8.

29. Molestina RE, Payne TM, Coppens I, Sinai AP (2003) Activation of NF-kappaB by Toxoplasma gondii correlates with increased expression of antiapoptotic genes and localization of phosphorylated IkappaB to the parasitophorous vacuole membrane. *J Cell Sci* 116(Pt 21):4359–71.

30. Barber GN (2001) Host defense, viruses and apoptosis. *Cell Death Differ* 8(2):113–126.

31. Ashida H, et al. (2011) Cell death and infection: A double-edged sword for host and pathogen survival. *J Cell Biol* 195(6):931–942.

32. Rudel T, Kepp O, Kozjak-Pavlovic V (2010) Interactions between bacterial pathogens and mitochondrial cell death pathways. *Nat Publ Gr* 8(10):693–705.

33. Sinai a P, et al. (2004) Mechanisms underlying the manipulation of host apoptotic pathways by Toxoplasma gondii. *Int J Parasitol* 34(3):381–91.

34. Nash PB, et al. (1998) Toxoplasma gondii-infected cells are resistant to multiple inducers of apoptosis. *J Immunol* 160(4):1824–30.

35. Youle RJ, Strasser A (2008) The BCL-2 protein family: opposing activities that mediate cell death. *Nat Rev Mol Cell Biol* 9(1):47–59.

36. Besteiro Sébastien (2015) *Toxoplasma* control of host apoptosis: the art of not biting too hard the hand that feeds you. *Microb Cell* 2(6):178.

37. Blader IJ, Koshy AA (2014) Toxoplasma gondii Development of Its Replicative Niche: in Its Host Cell and Beyond. *Eukaryot Cell* 13(8):965–976.

38. Coppens I, et al. (2006) Toxoplasma gondii sequesters lysosomes from mammalian hosts in the vacuolar space. *Cell* 125(2):261–274.

39. Sinai AP, Webster P, Joiner KA (1997) Association of host cell endoplasmic reticulum and mitochondria with the Toxoplasma gondii parasitophorous vacuole membrane: a high affinity interaction. *J Cell Sci* 110 (Pt 1:2117–2128.

40. Coppens I, Sinai AP, Joiner KA (2000) Toxoplasma gondii exploits host low-density lipoprotein receptor-mediated endocytosis for cholesterol acquisition. *J Cell Biol* 149(1):167–180.

41. Coppens I (2006) Contribution of host lipids to Toxoplasma pathogenesis. *Cell Microbiol* 8(1):1–9.

42. Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75(5):843–54.

43. Reinhart BJ, et al. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* 403(6772):901–906.

44. Pasquinelli AE, et al. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* 408(6808):86–89.

45. Fire A, et al. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* 391(6669):806–811.

46. Hannon GJ (2002) RNA interference. *Nature* 418(6894):244–251.

47. Elbashir SM, Lendeckel W, Tuschl T (2001) RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev* 15(2):188–200.

48. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T (2001) Identification of novel genes coding for small expressed RNAs. *Science* 294(5543):853–8.

49. Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science (80-)* 294(5543):858–862.

50. Lee RC, Ambros V (2001) An extensive class of small RNAs in Caenorhabditis elegans. *Science (80-)* 294(5543):862–864.

51. Bernstein E, Caudy AA, Hammond SM, Hannon GJ (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* 409(6818):363–366.

52. Hutvagner G (2001) A Cellular Function for the RNA-Interference Enzyme Dicer in the Maturation of the let-7 Small Temporal RNA. *Science (80-)* 293(5531):834–838.

53. Cai X, Hagedorn CH, Cullen BR (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* 10(12):1957–1966.

54. Lee Y, et al. (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* 23(20):4051–4060.

55. Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature* 432(7014):231–235.

56. Gregory RI, et al. (2004) The Microprocessor complex mediates the genesis of microRNAs. *Nature* 432(7014):235–240.

57. Yeom K-H, Lee Y, Han J, Suh MR, Kim VN (2006) Characterization of DGCR8/Pasha, the essential cofactor for Drosha in primary miRNA processing. *Nucleic Acids Res* 34(16):4622–4629.

58. Lee Y, et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature* 425(6956):415–419.

59. Yi R, Qin Y, Macara IG, Cullen BR (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev* 17(24):3011–3016.

60. Lund E, Güttinger S, Calado A, Dahlberg JE, Kutay U (2004) Nuclear export of microRNA precursors. *Science (80-)* 303(5654):95–98.

61. Khvorova A, Reynolds A, Jayasena SD (2003) Functional siRNAs and miRNAs

exhibit strand bias. *Cell* 115(2):209–216.

62. Winter J, Jung S, Keller S, Gregory RI, Diederichs S (2009) Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat Cell Biol* 11(3):228–234.
63. Britten RJ, Davidson EH (1969) Gene regulation for higher cells: a theory. *Science* 165(3891):349–57.
64. Wolter J (2013) "Gene Regulation for Higher Cells: A Theory" (1969), by Roy J. Britten and Eric H. Davidson. Available at: https://hpsrepository.asu.edu/handle/10776/6243 [Accessed September 24, 2017].
65. Ku H-Y, Lin H (2014) PIWI proteins and their interactors in piRNA biogenesis, germline development and gene expression. *Natl Sci Rev* 1(2):205–218.
66. Ng KW, et al. (2016) Piwi-interacting RNAs in cancer: emerging functions and clinical utility. *Mol Cancer* 15:5.
67. Matera AG, Terns RM, Terns MP (2007) Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol* 8(3):209–220.
68. Yu YT, Maroney PA, Darzynkiwicz E, Nilsen TW (1995) U6 snRNA function in nuclear pre-mRNA splicing: a phosphorothioate interference analysis of the U6 phosphate backbone. *RNA* 1(1):46–54.
69. Zhang Y, Cao X (2016) Long noncoding RNAs in innate immunity. *Cell Mol Immunol* 13(2):138–147.
70. Ramaprasad A, et al. (2015) Comprehensive Evaluation of Toxoplasma gondii VEG and Neospora caninum LIV Genomes with Tachyzoite Stage Transcriptome and Proteome Defines Novel Transcript Features. *PLoS One* 10(4):e0124473.
71. Ulitsky I, Bartel DP (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell* 154(1):26–46.
72. Liu SJ, et al. (2017) CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science (80-)* 355(6320). Available at: http://science.sciencemag.org/content/355/6320/eaah7111/tab-pdf [Accessed September 9, 2017].
73. Brennecke J, Hipfner DR, Stark A, Russell RB, Cohen SM (2003) bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell* 113(1):25–36.
74. Lai EC (2003) microRNAs: runts of the genome assert themselves. *Curr Biol* 13(23):R925-36.
75. Lagos-Quintana M, et al. (2002) Identification of tissue-specific microRNAs from mouse. *Curr Biol* 12(9):735–9.

76. Kim J, et al. (2004) Identification of many microRNAs that copurify with polyribosomes in mammalian neurons. *Proc Natl Acad Sci USA* 101(1):360–365.

77. Landgraf P, et al. (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* 129(7):1401–14.

78. Lagos-quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T (2003) New microRNAs from mouse and human. *Rna* 9(2):175–179.

79. Lim L, et al. (2003) The microRNAs of Caenorhabditis elegans. *Genes Dev* 17(8):991–1008.

80. Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 39(Database issue):D152-7.

81. Pall GS, Codony-Servat C, Byrne J, Ritchie L, Hamilton A (2007) Carbodiimide-mediated cross-linking of RNA to nylon membranes improves the detection of siRNA, miRNA and piRNA by northern blot. *Nucleic Acids Res* 35(8):e60.

82. Ramkissoon SH, Mainwaring LA, Sloand EM, Young NS, Kajigaya S (2006) Nonisotopic detection of microRNA using digoxigenin labeled RNA probes. *Mol Cell Probes* 20(1):1–4.

83. Válóczi A, et al. (2004) Sensitive and specific detection of microRNAs by northern blot analysis using LNA-modified oligonucleotide probes. *Nucleic Acids Res* 32(22):e175.

84. Kim SW, et al. (2010) A sensitive non-radioactive northern blot method to detect small RNAs. *Nucleic Acids Res* 38(7):e98–e98.

85. Chen C, et al. (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res* 33(20):e179.

86. Raymond CK, Roberts BS, Garrett-Engele P, Lim LP, Johnson JM (2005) Simple, quantitative primer-extension PCR assay for direct monitoring of microRNAs and short-interfering RNAs. *RNA* 11(11):1737–1744.

87. Varkonyi-Gasic E, Wu R, Wood M, Walton EF, Hellens RP (2007) Protocol: a highly sensitive RT-PCR method for detection and quantification of microRNAs. *Plant Methods* 3:12.

88. Brown KM, Blader IJ (2009) The role of DNA microarrays in Toxoplasma gondii research, the causative agent of ocular toxoplasmosis. *j ocul biol dis Inf* 2(4):214–222.

89. Chen Y, Gelfond J Al, Mcmanus LM, Shireman PK (2009) Reproducibility of quantitative RT-PCR array in miRNA expression profiling and comparison with microarray analysis. *BMC Genomics* 10(1):407.

90. Wang H, Ach R a, Curry B (2007) Direct and sensitive miRNA profiling from low-input total RNA. *RNA* 13(1):151–9.

91. Castoldi M, et al. (2006) A sensitive array for microRNA expression profiling (miChip) based on locked nucleic acids (LNA). *RNA* 12(5):913–920.

92. Nelson PT, et al. (2004) Microarray-based, high-throughput gene expression profiling of microRNAs. *Nat Methods* 1(2):155–161.

93. Schindelin J, et al. (2012) Fiji: an open-source platform for biological-image analysis. *Nat Methods* 9(7):676–82.

94. Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. *Nat Biotechnol* 26(10):1117–1124.

95. Margulies M, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376–380.

96. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589–595.

97. Mcelroy KE, Luciani F, Thomas T (2012) GemSIM: General, Error-Model based SIMulator of next-generation sequencing data. *BMC Genomics* 13(1):74.

98. Albers CA, et al. (2011) Dindel: accurate indel calls from short-read data. *Genome Res* 21(6):961–73.

99. Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* 11(1):31–46.

100. Hadfield J, Loman N Next Generation Genomics: World Map of High-throughput Sequencers. Available at: http://omicsmaps.com/.

101. Bentley DR, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456(7218):53–59.

102. Alessi J (Illumina) (2008) Illumina Genome Analyzer System.

103. Ma H-C (2011) Discovery and characterisation of new miRNAs during embryogenesis of D. melanogaster. Dissertation (University of Cambridge). doi:http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.609448.

104. Friedländer MR, et al. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 26(4):407–15.

105. Moxon S, et al. (2008) A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics* 24(19):2252–2253.

106. Braun L, et al. (2010) A complex small RNA repertoire is generated by a plant/fungal-like machinery and effected by a metazoan-like Argonaute in the single-cell human parasite Toxoplasma gondii. *PLoS Pathog* 6(5):e1000920.

107. Griffiths-Jones S (2013) miRBase blog: About. *miRBase*. Available at: http://www.mirbase.org/blog/about/ [Accessed August 22, 2016].

108. Tarver JE, Donoghue PCJ, Peterson KJ (2012) Do miRNAs have a deep evolutionary history? *BioEssays* 34(10):857–866.

109. Wang J, et al. (2012) A comparative study of small RNAs in Toxoplasma gondii of distinct genotypes. *Parasit Vectors* 5(1):1.

110. Braun L, et al. (2010) A Complex Small RNA Repertoire Is Generated by a Plant/Fungal-Like Machinery and Effected by a Metazoan-Like Argonaute in the Single-Cell Human Parasite Toxoplasma gondii. *PLoS Pathog* 6(5):e1000920.

111. XU MJ, et al. (2013) Comparative characterization of microRNA profiles of different genotypes of Toxoplasma gondii. *Parasitology* 140(9):1111–1118.

112. Hakimi M-A, Menard R (2010) Do apicomplexan parasites hijack the host cell microRNA pathway for their intracellular development? *F1000 Biol Rep* 2(42). doi:10.3410/B2-42.

113. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* 40(1):37–52.

114. Pavlakis GN, Jordan BR, Wurst RM, Vournakis JN (1979) Sequence and secondary structure of Drosophila melanogaster 5.8S and 2S rRNAs and of the processing site between them. *Nucleic Acids Res* 7(8):2213–38.

115. Wicks RJ (1986) RNA molecular weight determination by agarose gel electrophoresis using formaldehyde as denaturant: Comparison of rna and dna molecular weight markers. *Int J Biochem* 18(3):277–278.

116. Hafner M, et al. (2008) Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* 44(1):3–12.

117. Schröder J, Bailey J, Conway T, Zobel J (2010) Reference-free validation of short read data. *PLoS One* 5(9):e12681.

118. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.

119. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.

120. Saçar MD, Bağcı C, Allmer J (2014) Computational prediction of microRNAs from Toxoplasma gondii potentially regulating the hosts' gene expression. *Genomics Proteomics Bioinformatics* 12(5):228–38.

121. Xu MJ, et al. (2013) Comparative characterization of microRNA profiles of different genotypes of Toxoplasma gondii. *Parasitology* 140(9):1111–8.

122. Wang J, et al. (2012) A comparative study of small RNAs in Toxoplasma gondii of

distinct genotypes. *Parasit Vectors* 5(1):186.

123. Small Adapter Dimer Can Cause Big Troubles! - SEQanswers Available at: http://seqanswers.com/forums/showthread.php?t=22978.

124. Rubicon Genomics ThruPLEX Universal Low Input RNA LIbrary Prep Kit. Available at: http://rubicongenomics.com/applications/rna-seq/.

125. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17(1):10–12.

126. Kang W, FriedlÃ¤nder MR (2015) Computational Prediction of miRNA Genes from Small RNA Sequencing Data. *Front Bioeng Biotechnol* 3:7.

127. Hackl M, et al. (2012) Computational identification of microRNA gene loci and precursor microRNA sequences in CHO cell lines. *J Biotechnol* 158(3):151–5.

128. Lawless N, Foroushani ABK, McCabe MS, O'Farrelly C, Lynn DJ (2013) Next generation sequencing reveals the expression of a unique miRNA profile in response to a gram-positive bacterial infection. *PLoS One* 8(3):e57543.

129. Friedlaender MF, Mackowiak S (2012) miRDeep2 documentation version 2.0.0.5 [computer program].

130. Chiang HR, et al. (2010) Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev* 24(10):992–1009.

131. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics* 24(5):713–714.

132. Attention: Bowtie2 And Multiple Hits Available at: https://www.biostars.org/p/55237/.

133. de Hoon MJL, et al. (2010) Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. *Genome Res* 20(2):257–264.

134. Johnson NR, Yeoh JM, Coruh C, Axtell MJ (2016) Improved Placement of Multi-mapping Small RNAs. *G3 (Bethesda)* 6(7):2103–11.

135. ATTENTION: bowtie2 and multiple hits - SEQanswers Available at: http://seqanswers.com/forums/showthread.php?t=24270.

136. Castellano L, Stebbing J (2013) Deep sequencing of small RNAs identifies canonical and non-canonical miRNA and endogenous siRNAs in mammalian somatic tissues. *Nucleic Acids Res* 41(5):3339–3351.

137. Zeiner GM, Norman KL, Thomson JM, Hammond SM, Boothroyd JC (2010) Toxoplasma gondii infection specifically increases the levels of key host microRNAs. *PLoS One* 5(1):e8742.

138. Wiley M, et al. (2010) Toxoplasma gondii Activates Hypoxia-inducible Factor (HIF) by Stabilizing the HIF-1 Subunit via Type I Activin-like Receptor Kinase Receptor Signaling. *J Biol Chem* 285(35):26852–26860.

139. Cai Y, et al. (2013) STAT3-dependent transactivation of miRNA genes following Toxoplasma gondii infection in macrophage. *Parasit Vectors* 6(1):356.

140. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139–40.

141. Zhao J, et al. (2014) MicroRNAs expression profile in CCR6 $^{+}$ regulatory T cells. *PeerJ* 2:e575.

142. Dunay IR, et al. (2008) Gr1+ Inflammatory Monocytes Are Required for Mucosal Resistance to the Pathogen Toxoplasma gondii. *Immunity* 29(2):306–317.

143. Ho JJD, et al. (2012) Functional Importance of Dicer Protein in the Adaptive Cellular Response to Hypoxia. *J Biol Chem* 287(34):29003–29020.

144. Spear W, et al. (2006) The host cell transcription factor hypoxia-inducible factor 1 is required for Toxoplasma gondii growth and survival at physiological oxygen levels. *Cell Microbiol* 8(2):339–352.

145. Cleary MD, Singh U, Blader IJ, Brewer JL, Boothroyd JC (2002) Toxoplasma gondii asexual development: identification of developmentally regulated genes and distinct patterns of gene expression. *Eukaryot Cell* 1(3):329–340.

146. Feng X, Wang Z, Fillmore R, Xi Y (2014) MiR-200, a new star miRNA in human cancer. *Cancer Lett* 344(2):166–173.

147. Lin C-H, et al. (2009) Myc-regulated microRNAs attenuate embryonic stem cell differentiation. *EMBO J* 28(20):3157–3170.

148. Brunet J, et al. (2008) Toxoplasma gondii exploits UHRF1 and induces host cell cycle arrest at G2 to enable its proliferation. *Cell Microbiol* 10(4):908–920.

149. Molestina RE, El-Guendy N, Sinai AP (2008) Infection with Toxoplasma gondii results in dysregulation of the host cell cycle. *Cell Microbiol* 10(5):1153–1165.

150. Love MI, et al. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15(12):550.

151. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.

152. Robinson MD, et al. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11(3):R25.

153. Dillies M-A, et al. (2013) A comprehensive evaluation of normalization methods for

Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 14(6):671–83.

154. Tam S, Tsao M-S, McPherson JD (2015) Optimization of miRNA-seq data preprocessing. *Brief Bioinform* 16(6):950–63.

155. Fan Y, et al. (2016) miRNet - dissecting miRNA-target interactions and functional associations through network-based visual analysis. *Nucleic Acids Res* 44(W1):W135–W141.

156. Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* 33(Web Server issue):W741-8.

157. Wang J, Duncan D, Shi Z, Zhang B (2013) WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res* 41(Web Server issue):W77-83.

158. Andrews S FastQC A Quality Control tool for High Throughput Sequence Data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

159. Hi-Seq quality score behavior - SEQanswers Available at: http://seqanswers.com/forums/showthread.php?t=13155.

160. Nueda MJ, Tarazona S, Conesa A (2014) Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics* 30(18):2598–602.

161. Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG (2012) Functional microRNA targets in protein coding sequences. *Bioinformatics* 28(6):771–776.

162. Kanellos I, Dalamagas T, Hatzigeorgiou A, Fleming B Al, Athena RC (2014) MR-microT : A MapReduce-based MicroRNA Target Prediction Method. *SSDBM* (Aalborg).

163. Motenko H, Neuhauser SB, O'Keefe M, Richardson JE (2015) MouseMine: a new data warehouse for MGI. *Mamm Genome* 26(7–8):325–30.

164. Gao P, et al. (2009) c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism. *Nature* 458(7239):762–765.

165. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature* 460(7254):479.

166. Hwang H-W, Wentzel EA, Mendell JT (2009) Cell-cell contact globally activates microRNA biogenesis. *Proc Natl Acad Sci* 106(17):7016–7021.

167. Hong G, et al. (2014) Separate enrichment analysis of pathways for up- and downregulated genes. *J R Soc Interface* 11(92):20130950.

168. Kim S-W, et al. (2012) MicroRNAs miR-125a and miR-125b constitutively activate the NF-κB pathway by targeting the tumor necrosis factor alpha-induced protein 3

(TNFAIP3, A20). *Proc Natl Acad Sci U S A* 109(20):7865–70.

169. Hu S, Zhu W, Zhang L-F, Pei M, Liu M-F (2014) MicroRNA-155 broadly orchestrates inflammation-induced changes of microRNA expression in breast cancer. *Cell Res* 24(2):254–257.

170. Jiang S, et al. (2012) A novel miR-155/miR-143 cascade controls glycolysis by regulating *hexokinase 2* in breast cancer cells. *EMBO J* 31(8):1985–1998.

171. Cannella D, et al. (2014) miR-146a and miR-155 delineate a MicroRNA fingerprint associated with Toxoplasma persistence in the host brain. *Cell Rep* 6(5):928–37.

172. Rane S, et al. (2009) Downregulation of miR-199a derepresses hypoxia-inducible factor-1alpha and Sirtuin 1 and recapitulates hypoxia preconditioning in cardiac myocytes. *Circ Res* 104(7):879–86.

173. Wiley M, et al. (2010) Toxoplasma gondii activates hypoxia-inducible factor (HIF) by stabilizing the HIF-1alpha subunit via type I activin-like receptor kinase receptor signaling. *J Biol Chem* 285(35):26852–26860.

174. Gajria B, et al. (2007) ToxoDB: an integrated Toxoplasma gondii database resource. *Nucleic Acids Res* 36(Database):D553–D556.

175. Aurrecoechea C, et al. (2013) EuPathDB: the eukaryotic pathogen database. *Nucleic Acids Res* 41(Database issue):D684-91.

176. Krishna R, et al. (2015) A large-scale proteogenomics study of apicomplexan pathogens-Toxoplasma gondii and Neospora caninum. *Proteomics* 15(15):2618–28.

177. Blader IJ, Manger ID, Boothroyd JC (2001) Microarray analysis reveals previously unknown changes in Toxoplasma gondii-infected human cells. *J Biol Chem* 276(26):24223–31.

178. Okomo-Adhiambo M, Beattie C, Rink A (2006) cDNA microarray analysis of host-pathogen interactions in a porcine in vitro model for Toxoplasma gondii infection. *Infect Immun* 74(7):4254–4265.

179. PK13 ATCC ® CRL-6489™ Sus scrofa Available at: https://www.lgcstandards-atcc.org/Products/All/CRL-6489.aspx?geo_country=gb#generalinformation.

180. Nelson MM, et al. (2008) Modulation of the host cell proteome by the intracellular apicomplexan parasite Toxoplasma gondii. *Infect Immun* 76(2):828–44.

181. Kim S-K, Fouts AE, Boothroyd JC (2007) Toxoplasma gondii Dysregulates IFN- -Inducible Gene Expression in Human Fibroblasts: Insights from a Genome-Wide Transcriptional Profiling. *J Immunol* 178(8):5154–5165.

182. Saeij JPJ, et al. (2007) Toxoplasma co-opts host gene expression by injection of a polymorphic kinase homologue. *Nature* 445(7125):324–7.

183. Melo MB, et al. (2013) Transcriptional Analysis of Murine Macrophages Infected with Different Toxoplasma Strains Identifies Novel Regulation of Host Signaling Pathways. *PLoS Pathog* 9(12):e1003779.

184. He J-J, et al. (2016) Transcriptomic analysis of mouse liver reveals a potential hepato-enteric pathogenic mechanism in acute Toxoplasma gondii infection. *Parasit Vectors* 9(1):427.

185. He J-J, et al. (2016) Proteomic Profiling of Mouse Liver following Acute Toxoplasma gondii Infection. *PLoS One* 11(3):e0152022.

186. He J-J, et al. (2016) Transcriptional changes of mouse splenocyte organelle components following acute infection with Toxoplasma gondii. *Exp Parasitol* 167:7–16.

187. Illumina (2013) TruSeq Stranded mRNA Sample Preparation ®. (October):1–4.

188. Kim D, et al. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14(4):R36.

189. Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31(2):166–9.

190. Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37(1):1–13.

191. Timmons JA, et al. (2015) Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biol* 16(1):186.

192. Szklarczyk D, et al. (2017) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res* 45(D1):D362–D368.

193. Ganapathy V, Thangaraju M, Prasad PD (2008) Nutrient transporters in cancer: Relevance to Warburg hypothesis and beyond. *Pharmacol Ther* 121(1):29–40.

194. Singhal A, Jaiswal A, Arora VK, Prasad HK (2007) Modulation of Gamma Interferon Receptor 1 by Mycobacterium tuberculosis: a Potential Immune Response Evasive Mechanism. *Infect Immun* 75(5):2500–2510.

195. Chatterji U, et al. (2004) Indole-3-carbinol stimulates transcription of the interferon gamma receptor 1 gene and augments interferon responsiveness in human breast cancer cells. *Carcinogenesis* 25(7):1119–1128.

196. MA F, RD S (1993) The molecular cell biology of interferon-γ and its receptor. *Annu Rev Immunol* 11:571.

197. Rosowski EE, et al. (2012) Toxoplasma gondii Clonal Strains All Inhibit STAT1 Transcriptional Activity but Polymorphic Effectors Differentially Modulate IFNγ

Induced Gene Expression and STAT1 Phosphorylation. *PLoS One* 7(12):e51448.

198. Arzoine L, Zilberberg N, Ben-Romano R, Shoshan-Barmatz V (2009) Voltage-dependent anion channel 1-based peptides interact with hexokinase to prevent its anti-apoptotic activity. *J Biol Chem* 284(6):3946–3955.

199. Shoshan-Barmatz V, Zakar M, Rosenthal K, Abu-Hamad S (2009) Key regions of VDAC1 functioning in apoptosis induction and regulation by hexokinase. *BBA - Bioenerg* 1787(5):421–430.

200. Coppens I (2006) Contribution of host lipids to Toxoplasma pathogenesis. *Cell Microbiol* 8(1):1–9.

201. Barragan A, Sibley LD (2002) Transepithelial Migration of *Toxoplasma gondii* Is Linked to Parasite Motility and Virulence. *J Exp Med* 195(12):1625–1633.

202. Silva NM, et al. (2010) Toxoplasma gondii: The severity of toxoplasmic encephalitis in C57BL/6 mice is associated with increased ALCAM and VCAM-1 expression in the central nervous system and higher blood–brain barrier permeability. *Exp Parasitol* 126(2):167–177.

203. K??ppers M, Ittrich C, Faust D, Dietrich C (2010) The transcriptional programme of contact-inhibition. *J Cell Biochem* 110(5):1234–1243.

204. Kelman Z (1997) PCNA: structure, functions and interactions. *Oncogene* 14(6):629–640.

205. Fukami-Kobayashi J, Mitsui Y (1999) Overexpression of proliferating cell nuclear antigen in mammalian cells negates growth arrest by serum starvation and cell contact. *Japanese J Cancer Res* 90(3):286–293.

206. Mahon PC, Hirota K, Semenza GL (2001) FIH-1: a novel protein that interacts with HIF-1alpha and VHL to mediate repression of HIF-1 transcriptional activity. *Genes Dev* 15(20):2675–2686.

207. Metallo CM, et al. (2011) Reductive glutamine metabolism by IDH1 mediates lipogenesis under hypoxia. *Nature* 481(7381):380.

208. Fendt S-M, et al. (2013) Reductive glutamine metabolism is a function of the α-ketoglutarate to citrate ratio in cells. *Nat Commun* 4. doi:10.1038/ncomms3236.

209. Semenza GL, Roth PH, Fang HM, Wang GL (1994) Transcriptional regulation of genes encoding glycolytic enzymes by hypoxia-inducible factor 1. *J Biol Chem* 269(38):23757–63.

210. Ortiz-Barahona A, Villar D, Pescador N, Amigo J, del Peso L (2010) Genome-wide identification of hypoxia-inducible factor binding sites and target genes by a probabilistic model integrating transcription-profiling data and in silico binding site

prediction. *Nucleic Acids Res* 38(7):2332–2345.

211. Benita Y, et al. (2009) An integrative genomics approach identifies Hypoxia Inducible Factor-1 (HIF-1)-target genes that form the core response to hypoxia. *Nucleic Acids Res* 37(14):4587–602.

212. Shim H, et al. (1997) c-Myc transactivation of LDH-A: implications for tumor metabolism and growth. *Proc Natl Acad Sci U S A* 94(13):6658–63.

213. Kim J -w., Gao P, Liu Y-C, Semenza GL, Dang C V. (2007) Hypoxia-Inducible Factor 1 and Dysregulated c-Myc Cooperatively Induce Vascular Endothelial Growth Factor and Metabolic Switches Hexokinase 2 and Pyruvate Dehydrogenase Kinase 1. *Mol Cell Biol* 27(21):7381–7393.

214. Lee JY, et al. (2016) MCT4 as a potential therapeutic target for metastatic gastric cancer with peritoneal carcinomatosis. doi:10.18632/oncotarget.9523.

215. Young CD, et al. (2011) Modulation of Glucose Transporter 1 (GLUT1) Expression Levels Alters Mouse Mammary Tumor Cell Growth In Vitro and In Vivo. *PLoS One* 6(8):e23205.

216. Ullah MS, Davies AJ, Halestrap AP (2006) The Plasma Membrane Lactate Transporter MCT4, but Not MCT1, Is Up-regulated by Hypoxia through a HIF-1 -dependent Mechanism. *J Biol Chem* 281(14):9030–9037.

217. Chen C, Pore N, Behrooz A, Ismail-Beigi F, Maity A (2001) Regulation of glut1 mRNA by Hypoxia-inducible Factor-1: INTERACTION BETWEEN H-ras AND HYPOXIA. *J Biol Chem* 276(12):9519–9525.

218. Osthus RC, et al. (2000) Deregulation of glucose transporter 1 and glycolytic gene expression by c-Myc. *J Biol Chem* 275(29):21797–21800.

219. Schwartzenberg-Bar-Yoseph F, Armoni M, Karnieli E (2004) The tumor suppressor p53 down-regulates glucose transporters GLUT1 and GLUT4 gene expression. *Cancer Res* 64(7):2627–33.

220. Gan L, et al. (2016) Metabolic targeting of oncogene MYC by selective activation of the proton-coupled monocarboxylate family of transporters. *Oncogene* 35(23):3037–3048.

221. Mathupala SP, Ko YH, Pedersen PL (2006) Hexokinase II: cancer's double-edged sword acting as both facilitator and gatekeeper of malignancy when bound to mitochondria. *Oncogene* 25(34):4777–4786.

222. Bustamante E, Pedersen PL (1977) High aerobic glycolysis of rat hepatoma cells in culture: role of mitochondrial hexokinase. *Proc Natl Acad Sci USA* 74(9):3735–3739.

223. Rempel A, Mathupala SP, Griffin CA, Hawkins AL, Pedersen PL (1996) Glucose

catabolism in cancer cells: amplification of the gene encoding type II hexokinase. *Cancer Res* 56(11):2468–2471.

224. BOYLAND E, GOSS GCL, WILLIAMSASHMAN HG (1951) The hexokinase activity of animal tumours. *Biochem J* 49(3):321–325.

225. Rho M, et al. (2007) Expression of type 2 hexokinase and mitochondria-related genes in gastric carcinoma tissues and cell lines. *Anticancer Res* 27(1A):251–258.

226. Wolf A, et al. (2011) Hexokinase 2 is a key mediator of aerobic glycolysis and promotes tumor growth in human glioblastoma multiforme. *J Exp Med* 208(2):313–326.

227. Pastorino JG, Hoek JB (2008) Regulation of hexokinase binding to VDAC. *J Bioenerg Biomembr* 40(3):171–182.

228. Neary CL, Pastorino JG (2010) Nucleocytoplasmic shuttling of hexokinase II in a cancer cell. *Biochem Biophys Res Commun* 394(4):1075–1081.

229. Menendez MT, Teygong C, Wade K, Florimond C, Blader IJ (2015) siRNA Screening Identifies the Host Hexokinase 2 (HK2) Gene as an Important Hypoxia-Inducible Transcription Factor 1 (HIF-1) Target Gene in Toxoplasma gondii-Infected Cells. *MBio* 6(3):e00462.

230. Shulga N, Wilson-Smith R, Pastorino JG (2010) Sirtuin-3 deacetylation of cyclophilin D induces dissociation of hexokinase II from the mitochondria. *J Cell Sci* 123(Pt 6):894–902.

231. Finley LWS, et al. (2011) SIRT3 Opposes Reprogramming of Cancer Cell Metabolism through HIF1α Destabilization. *Cancer Cell* 19(3):416–428.

232. Alhazzazi TY, Kamarajan P, Verdin E, Kapila YL (2011) SIRT3 and cancer: tumor promoter or suppressor? *Biochim Biophys Acta* 1816(1):80–8.

233. Sundaresan NR, Samant SA, Pillai VB, Rajamohan SB, Gupta MP (2008) SIRT3 is a stress-responsive deacetylase in cardiomyocytes that protects cells from stress-mediated cell death by deacetylation of Ku70. *Mol Cell Biol* 28(20):6384–401.

234. Zhong L, et al. (2010) The Histone Deacetylase Sirt6 Regulates Glucose Homeostasis via Hif1α *Cell* 140(2):280–293.

235. Newsholme P, et al. (2003) Glutamine and glutamate as vital metabolites. *Brazilian J Med Biol Res* 36(2):153–163.

236. Aledo JC, Gómez-Fabre PM, Olalla L, Márquez J (2000) Identification of two human glutaminase loci and tissue-specific expression of the two related genes. *Mamm Genome* 11(12):1107–1110.

237. Szeliga M, et al. (2014) Silencing of GLS and overexpression of GLS2 genes cooperate

in decreasing the proliferation and viability of glioblastoma cells. *Tumour Biol* 35(3):1855–62.

238. Suzuki S, et al. (2010) Phosphate-activated glutaminase (GLS2), a p53-inducible regulator of glutamine metabolism and reactive oxygen species. *Proc Natl Acad Sci* 107(16):7461–7466.

239. Lee Y-Z, et al. (2014) Discovery of selective inhibitors of Glutaminase-2, which inhibit mTORC1, activate autophagy and inhibit proliferation in cancer cells. *Oncotarget* 5(15):6087–6101.

240. DeBerardinis RJ, et al. (2007) Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. *Proc Natl Acad Sci USA* 104(49):19345–19350.

241. Vander Heiden MG, Cantley LC, Thompson CB (2009) Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science (80-)* 324(5930):1029–1033.

242. Laborde E (2010) Glutathione transferases as mediators of signaling pathways involved in cell proliferation and cell death. *Cell Death Differ* 17(9):1373–1380.

243. Reitman ZJ, Yan H (2010) Isocitrate dehydrogenase 1 and 2 mutations in cancer: alterations at a crossroads of cellular metabolism. *J Natl Cancer Inst* 102(13):932–41.

244. Reitman ZJ, Parsons DW, Yan H (2010) IDH1 and IDH2: not your typical oncogenes. *Cancer Cell* 17(3):215–6.

245. Zhao S, et al. (2009) Glioma-derived mutations in IDH1 dominantly inhibit IDH1 catalytic activity and induce HIF-1alpha. *Science* 324(5924):261–5.

246. Robbins D, et al. (2012) Isocitrate dehydrogenase 1 is downregulated during early skin tumorigenesis which can be inhibited by overexpression of manganese superoxide dismutase. *Cancer Sci* 103(8):1429–33.

247. Wise DR, et al. (2008) Myc regulates a transcriptional program that stimulates mitochondrial glutaminolysis and leads to glutamine addiction. *Proc Natl Acad Sci USA* 105(48):18782–18787.

248. Gordan JD, et al. (2007) HIF and c-Myc: sibling rivals for control of cancer cell metabolism and proliferation. *Cancer Cell* 12(2):108–13.

249. Gordan JD, Bertout JA, Hu C-J, Diehl JA, Simon MC (2007) HIF-2alpha promotes hypoxic cell proliferation by enhancing c-myc transcriptional activity. *Cancer Cell* 11(4):335–47.

250. Koshiji M, et al. (2004) HIF-1alpha induces cell cycle arrest by functionally counteracting Myc. *EMBO J* 23(9):1949–56.

251. Doe MR, Ascano JM, Kaur M, Cole MD (2012) Myc posttranscriptionally induces HIF1 protein and target gene expression in normal and cancer cells. *Cancer Res* 72(4):949–57.

252. Li M, et al. (2002) Deubiquitination of p53 by HAUSP is an important pathway for p53 stabilization. *Nature* 416(6881):648–653.

253. Phan RT, Dalla-Favera R (2004) The BCL6 proto-oncogene suppresses p53 expression in germinal-centre B cells. *Nature* 432(7017):635–639.

254. Li M, Brooks CL, Kon N, Gu W (2004) A Dynamic Role of HAUSP in the p53-Mdm2 Pathway. *Mol Cell* 13(6):879–886.

255. Feng Z, Zhang C, Wu R, Hu W (2011) Tumor suppressor p53 meets microRNAs. *J Mol Cell Biol* 3(1):44–50.

256. Kim J-Y, et al. (2006) Toxoplasma gondii inhibits apoptosis in infected cells by caspase inactivation and NF-kappaB activation. *Yonsei Med J* 47(6):862–9.

257. Yang N, et al. (2013) Genetic basis for phenotypic differences between different Toxoplasma gondii type I strains. *BMC Genomics* 14(1):467.

258. Pavlides S, et al. (2009) The reverse Warburg effect: Aerobic glycolysis in cancer associated fibroblasts and the tumor stroma. *Cell Cycle* 8(23):3984–4001.

259. Bonuccelli G, et al. (2010) The reverse Warburg effect: Glycolysis inhibitors prevent the tumor promoting effects of caveolin-1 deficient cancer associated fibroblasts. *Cell Cycle* 9(August 2016):1960–1971.

260. Martinez-Outschoorn UE, et al. (2010) Oxidative stress in cancer associated fibroblasts drives tumor-stroma co-evolution: A new paradigm for understanding tumor metabolism, the field effect and genomic instability in cancer cells. *Cell Cycle* 9(16):3256–76.

261. Goldin N, et al. (2008) Methyl jasmonate binds to and detaches mitochondria-bound hexokinase. *Oncogene* 27(34):4636–4643.

262. Rotem R, et al. (2005) Jasmonates: novel anticancer agents acting directly and selectively on human cancer cell mitochondria. *Cancer Res* 65(5):1984–1993.

263. Fingrut O, Flescher E (2002) Plant stress hormones suppress the proliferation and induce apoptosis in human cancer cells. *Leukemia* 16(4):608–616.

264. Warburg O (1956) On the Origin of Gancer Cells. *Science (80-)* 123(3191). Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:On+the+origin+of+cancer+cells#0.

265. Crabtree HG (1929) Observations on the carbohydrate metabolism of tumours.

Biochem J 23(3):536–545.

266. YUSHOK WD (1958) Inhibition of glucolysis and fructolysis of Krebs 2 ascites carcinoma cells by chemical agents. *Cancer Res* 18(8 Part 2):379–389.

267. YUSHOK WD (1959) Metabolism of ascites tumor cells. I. Rate of glycolysis and competitive utilization of fructose, mannose, and glucose. *Cancer Res* 19(1):104–111.

268. YUSHOK WD (1964) METABOLISM OF ASCITES TUMOR CELLS. II. INHIBITION OF RESPIRATION BY GLYCOLYZABLE AND NONGLYCOLYZABLE SUGARS PHOSPHORYLATED BY HEXOKINASE. *Cancer Res* 24:187–192.

269. Yeung SJ, Pan J, Lee M-H (2008) Roles of p53, MYC and HIF-1 in regulating glycolysis - the seventh hallmark of cancer. *Cell Mol Life Sci* 65(24):3981–99.

270. Chiara F, et al. (2008) Hexokinase II detachment from mitochondria triggers apoptosis through the permeability transition pore independent of voltage-dependent anion channels. *PLoS One* 3(3):e1852.

271. Mathupala SP, Ko YH, Pedersen PL (2009) Hexokinase-2 bound to mitochondria: cancer's stygian link to the "Warburg Effect" and a pivotal target for effective therapy. *Semin Cancer Biol* 19(1):17–24.

272. Shulga N, Wilson-Smith R, Pastorino JG (2009) Hexokinase II detachment from the mitochondria potentiates cisplatin induced cytotoxicity through a caspase-2 dependent mechanism. *Cell Cycle* 8(20):3355–3364.

273. Agilent | Seahorse XF Instruments Overview and Selection Guide Available at: http://www.agilent.com/en-us/products/cell-analysis-(seahorse)/seahorse-xf-instruments [Accessed August 16, 2016].

274. DeBerardinis RJ, et al. (2007) Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. *Proc Natl Acad Sci USA* 104(49):19345–19350.

275. de Lima LPO, Seabra SH, Carneiro H, Barbosa HS (2015) Effect of 3-bromopyruvate and atovaquone on infection during in vitro interaction of Toxoplasma gondii and LLC-MK2 cells. *Antimicrob Agents Chemother* 59(9):5239–49.

276. Feldwisch-Drentrup H (2016) Candidate cancer drug suspected after death of three patients at an alternative medicine clinic. *Science (80-).* doi:10.1126/science.aah7192.

277. Som P, et al. (1980) A FluorinatedGlucose Analog , 2-fluoro-2-deoxy-D-glucose (F-18): Nontoxic Tracer for RapidTumor Detection ofacute orchronic. *J Nucl Med* 21:670–675.

278. Smith AB, Smirniotopoulos JG, Rushing EJ (2008) From the archives of the AFIP:

central nervous system infections associated with human immunodeficiency virus infection: radiologic-pathologic correlation. *Radiographics* 28(7):2033–2058.

279. Hoffman JM, et al. (1993) FDG-PET in differentiating lymphoma from nonmalignant central nervous system lesions in patients with AIDS. *J Nucl Med* 34(4):567–575.

280. Sandherr M, et al. (2001) Pitfalls in imaging Hodgkin's disease with computed tomography and positron emission tomography using fluorine-18-fluorodeoxyglucose. *Ann Oncol* 12(5):719–722.

281. Garrison MA, Glanton C, Rasnke M, Smith ME, Ornstein DL (2002) Challenging cases and diagnostic dilemmas: case 1. Tracheal compression in Hodgkin's disease. *J Clin Oncol* 20(15):3344–3347.

282. Slater GSC, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.

283. Warburg O über den Stoffwechsel der Carcinomzelle. Available at: http://www.springerlink.com/index/N167Q8W9V0348PR8.pdf.

IX. Appendices

9.1 Putative novel *Mus musculus* miRNAs from 3.4.1

9.2 Putative novel *Toxoplasma gondii* miRNAs from 3.4.2

9.3 Putative novel *Mus musculus* miRNAs from 5.3.1

9.4 Gene lists from 6.3 are included in an external DVD, as Supplementary Materials

9.1 Putative novel *M. musculus* miRNAs, from 3.4.1

provisional id	miRDeep2 score	total read count	mature read count	loop read count	star read count	consensus mature sequence	consensus star sequence
1_1152	68.1	181	180	0	1	uaguuuuacauuuuuuauuuu	aaauaaaaaauaauaauaua
10_2483	27.1	66	65	0	1	uuuuacagucucauaaua	auuaugaauacugcauuu
11_2824	9.9	25	21	0	4	uuguauucucccuuccuu	ugggggggggguuguuuuuuu
17_8492	9.9	18	14	0	4	acucucucacucugcaugguuac	ugacuuucguacggagagagagagaaa
16_8372	5.2	35	35	0	0	augugugugguauguguguau	acacacauacauacacacauaa
2_11921	5.2	192	192	0	0	ugagguagaaggcugugugc	acacgccugcacucucggc
12_3883	4.8	61	61	0	0	gaggguuggguggaggcug	gcaccaccaugcccagcu
X_20178	4.7	19	19	0	0	gggggugcagcucagugg	acuuuccucacauaccccag
14_6561	4.5	12	12	0	0	uguuauagauuccugcuc	gugggaaacuuuggacauu
10_1912	4.4	230	230	0	0	uuguacagugugaugcuc	gcucugcuucugaag
9_19150	4.3	10	10	0	0	uggcaguggaguuagugauugu	aaucagcuaauuacacugccuac
14_6034	4.3	17	17	0	0	agcagcauuccacagguc	caaagguggagggcugugaa
11_3513	4.1	13	13	0	0	ugagguaguugugugguu	cuacauaccuagccucaua
16_8320	4	1095	1095	0	0	uguuauagugugaggaga	accauggcuuuaugacauc
9_19433	4	14	14	0	0	uacauucauugcucucug	augaguggagaaaguaag
10_2617	3.6	816	816	0	0	ugagguaguggguuguuu	uuaaccuccuccauuaga
5_14886	3	548	548	0	0	uguuauagagugaguaua	uucucacuaaugugauagu
1_1519	2.9	9	9	0	0	cuauacagguguucuuccuuu	cuagagagcuagcugucuuaguu
6_15778	2.1	89	89	0	0	uucuccagucccgccguc	cggcguggacguggugcu
5_15035	1.9	39	39	0	0	uucuauggccuguucuuc	ggaauuggcuagcagggc
2_10828	1.9	20	20	0	0	ugagguaguauuuuguuu	gcaaaaugcugguucuggu
7_17481	1.9	30	30	0	0	uguuauagucucauaauc	ugauguagauuugaaaggauggu

7_17484	1.8	260	260	0	0	gggggguguagcucagugguaga	agccccugccuagaaucccccag
13_4913	1.8	20	20	0	0	uucuacaucccucccccuc	gggggggggagcguguuggg
4_14431	1.7	551	551	0	0	uucuaaagcccaccgucc	cuggugggaggagug
10_1721	1.7	21	21	0	0	cugugcucuccauuccuc	ggaagcggagccgcacagug
2_11195	1.7	144	144	0	0	cgcuacaguccgccgagc	ugggcggcucuguggguggc
5_15644	1.6	424	424	0	0	agcucuacagucagaagcuc	gcucagcugugcagucag
13_5378	1.6	195	195	0	0	uuuuauagucugaggcuc	gccugagucuguggcg
7_17091	1.6	1073	1073	0	0	uucuauagugugaggaga	uccuuguguaggaca
12_4157	1.6	115	115	0	0	uucuacauucccuccugc	aggaguggcaggc
7_17622	1.6	13	13	0	0	ugauauagccaagcccgacugua	cagucaggcugcuggcuauauuccagg
17_8931	1.6	37	36	1	0	uucuccagccccccauuc	uuuaggggugggagaagu
8_18839	1.5	156	151	5	0	uuuuuuaaguucuaagauu	ucccagcacucgggaagca
14_6481	1.5	22	22	0	0	ggggauguagcucaguggcagag	cagcccuagguuggaucccacg
11_3196	1.5	33	33	0	0	uuugcaguaacaggugguggacaucc	guggucauguaugauacugcaaaca
11_3031	1.5	84	84	0	0	uccuacacccccccagc	ugggggggggggaag
11_3226	1.5	97	97	0	0	uucagggauaaaauggagucacaga	ugugacuccugagcucuguucc
3_12951	1.5	16	15	0	1	aguaccaccauacacagcuuuugu	gagcuggguguggguggcacaugc
15_7353	1.4	18	18	0	0	uagagcuguacagucaua	uugcauguaucagcucuugc
5_15643	1.4	430	424	0	6	agcucuacagucagaagcuc	agcuguacagucagaagc
11_3385	1.4	51	51	0	0	gggggguguagcucagcggua	cucuguguuauuugaucccccag
19_9928	1.4	94	94	0	0	gggggguguagcucagugg	ccugaguucaaccac
17_8647	1.4	126	126	0	0	aggggguggggaaaaaaaa	ggacuuuccacccuuagc
3_13243	1.3	50	50	0	0	uucuaaggcccaccauuc	uuauuggguuuugaacc
14_6524	1.3	931	931	0	0	uucuccagucccacccuc	ggaaggcuggagccag
10_2398	1.3	513	513	0	0	aguuauagagucauaaga	uuauguucugacugu
19_10198	1.3	1276	1276	0	0	agggagggaacgcagucugagugga	cauugaugaucguucuucucuccuucg

12_4373	1.3	34	34	0	0	uguuguaguagccuguguuc	acugucaaacagguacugugacaau
16_8393	1.3	25	25	0	0	uucugcuguccugugcuc	gccagggcccuauggauug
9_19001	1.3	30	30	0	0	uucugcagucugcaguuc	ucugugacccaacugcaggauu
X_20698	1.3	16	16	0	0	agggggaggcaaaaaaaaa	ccuuucugccuccucugcc
1_1253	1.3	43	43	0	0	ucagagagcuacaggucc	acccagucucaaaggug
7_17963	1.3	2511	2511	0	0	uucuacagugugaggauc	gccucgccuuagaccagaguc
1_715	1.3	91	91	0	0	uccuacugcuccaugcuc	gcaggggcagguuccau
16_8088	1.3	30	30	0	0	uuuuccagucccacccuc	gggggagaaugagagacc
6_16407	1.2	146	146	0	0	uucuaaagaccgaagagc	ucagcgguuaaaag
15_7116	1.2	325	325	0	0	uucuucaggcccaccauc	uguuggggguugaggggga
4_14426	1.2	12	12	0	0	auacagucuccaacaauc	uuguuggaggaaguguauuu
4_13891	1.2	732	732	0	0	uucuacagcccgaccucc	aggaguggugcaggggg
5_15620	1.2	238	238	0	0	ugaggcaguaccuuguuc	acaguguauggcuugcagg
1_716	1.1	91	91	0	0	uccuacugcuccaugcuc	gcccgugggucuucuggaag
10_1763	1.1	52	52	0	0	ugucauacagucccccgcuc	guacuggacugguugacaga
3_12945	1.1	29	29	0	0	ucguccucucccuuccuc	ggggggggggaagaucagu
17_8834	1.1	11	11	0	0	ggaaggugggugcuaagggcuga	aguucauagcaacccccccuuaagccug
2_11126	1.1	21	21	0	0	uucuccauucccaugcuc	gcuugagaguggacagca
5_15110	1.1	16	16	0	0	ugagugugugugugauuguga	auuuuggaugugcauacaugugug
1_1307	1.1	153	153	0	0	uucuauaguguuauuauc	aaguaauacuagaagaacu
X_20633	1.1	20	20	0	0	ugcuaucuuccuuccuuc	gggcaggaagcagg
10_2577	1.1	115	115	0	0	ugcuguagucuguuguuc	augacagccuacuccaugcaac
9_19633	1.1	39	39	0	0	uaguguucuccaucccuc	gggguggauuugau
1_1544	1.1	100	100	0	0	aguuauacaguuauaugaua	ucauauacauauuugu
11_3633	1	393	393	0	0	aucucuccagucccacucuc	gagugaggcugggacccugg
11_2654	1	271	271	0	0	ugaggcaggagguugugu	ccagccaaugucaccaauc

1_310	1	11	11	0	0	cuguauccucccuuccgc	gggagucagagggucucagag
7_17974	1	120	120	0	0	uucugugcucucuugucc	auaggggcaggacaga
3_12577	1	81	81	0	0	uccuacucucccuuccuc	gacagggacuggggcuuggagg
10_1874	1	12	12	0	0	ugagggggagagagggug	ccacaguuuccccucugu
2_11512	1	58	58	0	0	uccuacugcuccaugucc	acuugcuugccaguacaaggauu
5_15253	1	230	230	0	0	uguuauacaguacgauga	auugugcuaggacaca
2_10551	1	133	133	0	0	ugcuauagugugagaauc	gucuugacucaaauauaagauca
10_2048	0.9	32	32	0	0	agggggugggggguuugg	agcaaccucaucaacggg
12_3832	0.9	27	27	0	0	ucguauugcuccauuucc	aaaugggcagggaga
9_19209	0.9	1713	1713	0	0	uucuacagccccuccauc	uggaaaggaaaaugucaugaaga
18_9407	0.9	19	19	0	0	cucuguucuccauuccuc	gagagagagacagagag
X_20069	0.9	47	47	0	0	ugaggcaggaggauguac	acaugccuguaagc
14_6021	0.9	29	29	0	0	acccagucuccaccaauc	ucacuggggcuuugagguca
4_13560	0.9	2165	2165	0	0	uuuuauagugugaggauu	gccuugccuauggagga
8_18786	0.8	158	158	0	0	uucuccaguccuccccuc	ugugguaggaacaggaagg
5_14733	0.8	12	12	0	0	uucuauguucucccuuuc	ggcaugaaggaga
15_6865	0.8	97	97	0	0	uuguauagcuuguuguuc	acugugcaaguaugcuaug
5_15628	0.8	53	53	0	0	uucuacagagccauaaaua	aaaugugcucuguagcaca
8_18072	0.8	6566	6566	0	0	uuuuauagugugaugaga	ugauuuguuuucacuguaacccu
2_11711	0.8	1240	1240	0	0	uucuacccucccuccuuc	gggggggggggguaguggu
3_13001	0.7	12	12	0	0	uucuacacccucuuccuc	ggacgagaggugugauau
X_20208	0.7	11	11	0	0	uccuagaucuccauuauc	caguggaggcugagguag
6_16742	0.7	11	11	0	0	ucagacagcuccaccccc	ggguggaggucagcca
7_16877	0.7	75	75	0	0	aguucuacauccuacucuc	ggauggaggugcuagaguca
6_16639	0.7	594	594	0	0	ugcuacagacccaggauc	uccugaucugugcaug
2_11768	0.7	785	785	0	0	uucuaccaucccaccauc	uggucugcaggaugggagaaga

16_7896	0.7	13	13	0	0	agaguucuacauccgauc	cuggaagcagagccucaaa
5_15531	0.6	25	25	0	0	ugcuacauucucuccuug	gggagguggagauggcagu
3_12752	0.6	41	41	0	0	auggagauagauauagauau	guguauauauauguacauau
7_17365	0.6	52	52	0	0	uguuauauaguuauuaua	auguaacuguauagcuau
7_17874	0.6	17	17	0	0	ugagguaggggcgguguuc	uccugcguuccaaccacagc
15_6958	0.6	9	9	0	0	aacugaguugaaggcaaaggua	gcuuugccuucagcuuagucgu
18_9202	0.5	215	215	0	0	uuauacauucagaauauc	uaucuugggugcucagca
11_2984	0.5	100	100	0	0	uuguacacuccauccuuc	agugaucaugugcgcauug
4_13571	0.5	80	80	0	0	uucucuaguccuccuauc	uauggggcuaguuauca
14_6335	0.5	234	234	0	0	uuguauucuacuuuccuc	ggaaagcuguggaccagcc
19_10124	0.5	18	18	0	0	cagaagguuggccauuggggaa	cccaaauggcucacccucuccu
1_1075	0.5	875	875	0	0	uucuaaaaucccacgauc	auggggagauagaagg
13_5038	0.4	83	83	0	0	uaguucuacauuuugauc	aggaaggaagaacaauc
17_8704	0.4	128	128	0	0	auaggugugugguauguguguau	acaguguauaggggugugugugc
15_7325	0.4	1447	1447	0	0	uucuacagucccucccuc	ggccucccucuguggaguu
12_3810	0.4	265	265	0	0	uucuauagagucagaaua	uucugccuuagucuu
1_1004	0.4	23	23	0	0	gggggugcagcucagugg	auuggagauuaaccaucccag
13_4924	0.4	58	58	0	0	uuuuauauugugcuuaua	uggguaaauguguacagagua
13_5389	0.3	83	83	0	0	uucucuacucucccaauc	uugggagcugagaguuug
3_12426	0.3	230	230	0	0	uuguacagugugaugcuc	gcucugcuucugagg
1_242	0.3	36	36	0	0	auagaugugugugcaugugguguau	gugcguaugugugugugucuguau
19_10042	0.3	801	801	0	0	ugaggcaggagguuguuu	accagccuccuccuugcu
14_6308	0.3	1424	1424	0	0	uucuacacucccgccauc	ugugaggagaugagaagc
1_825	0.3	71	71	0	0	ugaggcaguacauuguac	acgaccuguacuguccuugag
6_15816	0.2	170	168	1	1	ucuuuucaguucuucuauu	agagggagagagagagagagagaga
18_9408	0.2	19	19	0	0	cucuguucuccauuccuc	ggauuggaaaacagcaac

17_9030	0.2	58	58	0	0	ucccugagacccuuuaac	cagaggugagggaga
8_18550	0.2	97	97	0	0	uucagggauaaauggagucacaga	ugugacuccugagcucuguucc
10_2550	0.2	4401	4401	0	0	uucuacagucccucguuc	augauguccuguguaug
17_8435	0.2	8	8	0	0	uuauacauuuuuauucuu	gauuaaaggcgugugcca
8_18552	0.2	97	97	0	0	uucagggauaaauggagucacaga	ugugacuccugagcucuguucc
X_20518	0.2	109	109	0	0	ugcuccagacccacacgc	gugugggacacuggaacugc
5_14769	0.2	5	5	0	0	caccacagucccacgagc	acguggcucuuugcugau
10_2238	0.2	87	87	0	0	uuagacaguccauuguuu	acaccugcagagu
5_14785	0.2	1930	1930	0	0	uucuccaguccgacucuc	gagucaaggcugugugacuc
6_16084	0.1	67	67	0	0	uuguacucuacauugagc	ucacuguugagaugucu
18_9405	0.1	11	11	0	0	uugagcuguacagugauc	uccugccuggcuuguga
14_6353	0.1	463	463	0	0	aucuaccuuccgacgauc	uuucuagcgaagaggc
15_7030	0.1	80	80	0	0	uucuacucuaccuccauc	ucagggugagaugcuaguaauu
6_16174	0.1	305	305	0	0	uucuauucucccuuccuc	gguagggugcuuagaaau
1_1211	0	44	44	0	0	uucuauaguguuaggagc	ugaaaauucuggaugaaa
1_1543	0	100	100	0	0	aguuauacaguuauaugaua	ucauauacauauuuguauaguaac
7_17551	0	14	14	0	0	caugugugcauugaggaucuu	gaucuguugcaaugggucacuggu
14_6234	0	27	27	0	0	gggggugcagcucagugg	accuggaguuugaucucuag
6_15801	0	208	208	0	0	ugaggcaggcggaugauu	uacugcuuaaauguucuagu
11_2688	0	205	205	0	0	ucauauugcuccauguuc	acauacuguaaugugauu
18_9267	0	474	474	0	0	uuauauaguagguuguuu	acaaccaagaauauauaaag
3_12530	0	2440	2440	0	0	uucuacaguacuaccauc	ugggcacugacuguaggccu
X_20810	0	14	14	0	0	uuuucuauagugauuaua	gaaucacgaggaaaug
4_14417	0	230	230	0	0	uuguacagugugaugcuc	gcucugcuucugagg
9_19567	0	146	146	0	0	uucuauucuccuuuccuc	gguggggaauggaggc
10_2191	0	26	26	0	0	ucuuuucccuccuucuauu	ucugaggaauuaaagaug

9.2 Putative novel *T. gondii* miRNAs from 3.4.2

provisional id	miRDeep2 score	total read count	mature read count	loop read count	star read count	consensus mature sequence	consensus star sequence
TGME49_chrVIII_2086	2.2	36	36	0	0	uucuccauucccacacuc	uagugggauuggagggcc
TGME49_chrII_3466	1.6	16	16	0	0	uuguguuguacauuguuc	augaagucaacaagauc
TGME49_chrX_1032	1.5	11	11	0	0	ugcgacagacagucgauc	ucguuuugugucugcauu
TGME49_chrXII_519	1.2	10	10	0	0	aguucuucagucugccgagu	acgagcagaagaacgaacugg
TGME49_chrIV_3185	1.1	114	114	0	0	uucuccacaccgcugcuc	gcggccuggaguccu
TGME49_chrVI_2926	0.2	85	85	0	0	ucauauaucuacauguuc	ccauguacguauauauau

9.3 Putative novel *M. musculus* miRNAs, from 5.33

provisional id	miRDeep2 score	total read count	mature read count	loop read count	star read count	consensus mature sequence	consensus star sequence
13_9745	33140.4	65011	65010	0	1	cugggcuacacauuuuua	aaaaagcagagguagagcc
3_28811	5118.3	10056	10054	0	2	ucccggguuucggcacca	gaguuggaauuuugauggggauu
7_39899	3695.4	7251	7244	4	3	ugagcugucuccuaguaccuauu	aaaggacuagagacuaaa
11_3926	1703.7	3373	3361	6	6	ugguggguгcauauguuu	aacucuaccuuccuaauuuu
14_11408	1406.2	2758	2680	1	77	gaccgaguaacugcuagauccu	agucuagcucuugcucuugcuc
11_3345	1401.6	2757	2680	0	77	gaccgaguaacugcuagauccu	agucuagcucuugcucuugcuc
13_9832	1310.8	2564	1696	0	868	uagguagaccaggcugaucu	uguagcccuggcuguccugga
2_25682	1204.4	2367	2358	0	9	ucccgcggcuccugcucugcuc	agcgaggggcggagagccgcgc
9_44382	939.5	1836	1797	2	37	uagguagaccaggcugaucu	auccuccugucucugccuucu
3_29821	922.2	1818	1816	0	2	cuccauguaucuuugggaccugcc	cagucucccuuccuagccaugg
7_38813	645.1	1264	1255	0	9	aaagagagacagauagauagag	cuaucuaucuaucuaucuaucugu
8_42888	642.8	1266	1251	0	15	uguccgggggaccgacuuugcc	auggagccgcgcuccgggacgcg
4_31937	444.1	876	874	0	2	ucgcgggcugccgagcuccaggucc	ccgggucccggcagcugcggccc
10_1343	388.3	776	766	0	10	uuggcugagaaaaaugacugaauag	uauucauauacuucuagccuaca
3_29313	362.5	709	343	0	366	uccgagcuccgagccccgaggcag	cccuggggcucuggagcccgaga
17_16347	351.5	681	648	0	33	uuggucugagcaucuuccagg	gggaggaugucaggaugcagacug
17_17490	351.4	681	648	0	33	uuggucugagcaucuuccagg	gggaggaugucaggaugcagacug
12_7670	331.1	648	646	0	2	uuaggaaugcuaggcuagugcu	acuagcuugcuuuccugggg
17_17246	306.1	591	555	0	36	aggccugcucugagccuccgccu	gagggcccagagcagagguucu
9_43646	290.3	561	560	0	1	uggcaguggaguuagugauugu	aucagcuaauuacacugccuaca
7_38735	280.6	550	547	0	3	ucugucuucucuugugccuauuc	guaagcacaggaagccacaggcu
15_12114	269.9	527	526	0	1	aacugaguugaaggcaaaggu	gcuuugccuucagcuuagucgu

8_42936	266	513	474	4	35	aacucagaucugccugccucug	aggcaagaaguucugagcuaga
12_6448	247.6	484	451	0	33	ucagaacaaccugaccugccu	ccagcacugaguuguucuguca
10_1853	237.1	474	375	0	99	caugacugaaacucgacauccg	agagucgggucucagccagucu
12_7958	236.5	462	427	0	35	ucugaucccaacccuucugcccagc	agggcagaagugagguuaggagc
16_14002	228.5	455	412	0	43	agccaugacggaagacuguguu	gacuguguucuguguuggugc
3_29798	228.1	446	435	0	11	ucugagccccgagacuguauuac	uuacaguccgagcccugagacu
14_11592	219.6	429	413	0	16	ucagccacuguguccccuccu	cuuggggcagagguggcacggugu
10_6	218.1	426	249	0	177	gagggacauacucaaugagaac	ucuuauuguccauguccugcc
7_40090	215	428	371	4	53	acugccuccugccugcccugcaga	ugacagggugggugggcagccc
2_25278	210.7	413	412	0	1	aacaccaggacugaaaacagccu	uuguuucaucuccuggguuugu
2_24316	204.6	400	303	0	97	uaaggcaugcaugcuucaggcu	ucugugucaugcauuccacacu
4_31673	200.8	392	388	0	4	uucaaaccucucuggcugcc	caggccaaggugggauuuuugaggg
6_36022	194.1	380	375	0	5	uuauggcccuucgguaauucacug	ugaauucuaccagugccauaca
11_4468	188.4	368	285	0	83	ucuguuggauccugugaggaca	cccauaggauuuuaacagaacug
11_3915	184.2	367	365	0	2	ugcacccuucugacccacuucuccu	gaaggaggugggggcugcugug
9_43534	181.3	354	304	0	50	ccaggcugcuggagucugggu	ccagucuccaucugccuuccu
4_31706	174.3	340	335	2	3	acuuucuagauguucccggaag	cugggaacaucuacaaagugaagu
X_46099	168.6	329	317	0	12	cugagcacccuggccccccaga	aaggggaucagggauccaga
19_19896	167.3	336	234	0	102	ugggcuccgccugugguccgc	cugacuuccaggcccagcccugc
17_16661	155.8	315	293	0	22	ugugggaugaauggcaccugg	aggccuucuuucugucccaccca
11_2468	155.6	315	293	0	22	ugugggaugaauggcaccugg	aggccuucuuucugucccaccca
2_26011	155.5	297	251	0	46	auauacauuaaauaagucaaug	uugauuuauuuaauguauaug
17_16779	149.2	297	294	1	2	cgggggggccggcggcggcg	ugacuggcggcgcccgcgcagc
8_41659	140.7	275	264	0	11	ggucccugauaucgaugcugugc	ucggcaccgacacaaggauccug
10_690	132.2	258	230	0	28	uuugguuccucugaccuuuugcu	gccaggucucuggagccuuug
3_29076	131.3	256	255	0	1	ucuguggguguguuuguccgucc	aaggagacacaguccuugcaggc

10_711	124.6	251	244	0	7	augaccuggccuugcucauc	guggguaagcugugguccuggccugu
15_13237	122	238	88	0	150	cgaggcaucucacaggccgucu	uugggcugucagagagaugg
13_9699	111.1	218	199	0	19	cacucugaaaaugcagaugagcu	uucauuguugauuuucgguggga
10_1972	106.4	208	206	0	2	caagcuggacuccaggccccaga	ugggcuuggauagccagccacgcc
6_37662	103.5	192	182	0	10	ugacuggcagaggaagcucacc	gugagcuuccucugccaguc
7_40779	101.3	198	171	0	27	acucacucuguagaccag	uggccucagacugagagaucu
16_15096	91.1	176	151	0	25	ugccuguggaagucagaugagg	cucuucuggccuucaugggcacu
8_40983	89.9	183	181	0	2	uugcucugugcuguggaucaggagc	ucccugagccucuggagagca
1_23755	88	171	148	1	22	ucugaccccuggacucacugga	agugacucucuaaggcaggacc
4_30295	87.1	169	156	0	13	gaggaucugguucuguagcuc	ugcuacaggaugagauccggucu
3_29436	78.6	145	143	0	2	ugagcaccuccugcaccccaugg	cuggggggugaggugggg
2_24211	77.3	150	64	6	80	cuccaaagcucccagaggcuagg	ugccucuggaagguuuggcagg
11_3180	76.7	149	87	1	61	acagucugucaccugagccaaacu	uuugcugcauuugacaggcacag
2_24248	76.5	149	131	0	18	ucucuaucccuggccugcucuc	gccuaggguuuggggauacagagc
11_6018	75.7	146	138	0	8	ucagccaugcuaaggccugc	ugggcuggggcguggcucagc
2_25073	75.5	159	151	0	8	acgugaccauuucugcccaauu	cuggggagaauggccauuuca
5_35384	75.4	154	152	0	2	ugccuaggugauagcggaucgcg	agagccguuaucgguugggccuc
11_3502	75	145	127	12	6	acugggcugcucugggcgagccgg	cugcacaccuggagcccagauaa
16_15400	72.3	140	138	0	2	uucugaucucuucccucccuccu	gagggaggggaaagauaagaaaa
17_17068	71.9	149	145	0	4	ugucuggaucugacccacaga	uggugggcagagcucagucaguggg
11_5359	68.2	132	64	0	68	gugggcaugcaguaguggaccagc	agcuccacuguugucccccacagc
5_34825	67.8	141	79	9	53	uaugaguucuaggcauugaaucu	gucaaugcucugaacuccuaag
15_13680	66.1	128	109	0	19	ccgaugccauucccccgccgccc	aggcgggggguggcaaugc
3_28397	62.9	127	126	0	1	ucugucauuuuucccaccu	agugggaaaauugacauggauugu
2_27003	62.3	122	116	0	6	ugucuguacaguucaggugggga	ucaaccccaacuguccagucgcu
8_42162	54.2	107	92	0	15	aaacugucugucuguucauagc	uaugagcagaccuccaguuuau

7_40715	53.9	117	113	0	4	uuggggaaggcaguacuuaugu	auggguguccuccuucccaga
10_600	53.4	103	90	0	13	acccaggaagcuggggcaggag	cuaccucagucuccuugagcgc
12_8133	53.4	103	95	0	8	uauguggauucucucacuagcc	cugugggagagucuccagacu
9_43572	53.2	97	39	0	58	cgcuacauugauuggacacuga	aguggacaguucuguguggccu
10_598	51.2	99	97	0	2	uaaacagaccaggcuggccuc	gaucagcaugucucugcuucuc
7_40058	48.1	100	81	0	19	acugguacaaggguuggggagac	ucccccaggccuguaccaggguc
11_2710	47.2	93	92	0	1	ucugauccugugcaguacugga	caguacuggucaggaucacc
17_15871	47.1	90	58	0	32	cgggcacggggacacugacgug	cccucaguguucccguguggcc
2_24119	45.6	87	84	0	3	ugcaaugccuggaucugggcuc	uucccaggcuuaggcauugcaca
11_4107	44.2	85	67	0	18	uccugcccccuuucccuguaga	uaaggguggaugggggcagagcu
12_6962	43.7	92	84	0	8	cuuccuccuugacugggucauc	agcacccggagaaggagggcggac
7_38181	42.9	83	79	0	4	ucuucucuuuccagucaucagc	uggugccuggauuggaggaugaga
17_15524	42.9	82	80	0	2	cuuccaagugcuggacuacaggc	cugagggcaggacuuggagc
15_12512	42.7	93	71	0	22	uuccagaucuacaagccagagcuu	cccuagcuaguauaucuugaaucu
9_44170	41.2	86	64	0	22	cgccgggguuggcccugagc	cucucggccccgccccggcccu
19_20413	40.5	91	28	19	44	guaagggaccagggcaggagga	cgcucacccuucuccuuuccu
2_25292	40.4	77	76	0	1	aauuacccaccaaaccccagaag	uuuuggggggggggggguauugu
12_7318	39.7	84	28	0	56	cuuccuaccuuuuccuggcgg	cccuaggaccagguaggaguugc
1_22957	39.1	76	40	0	36	caaagcguuuuaagcggaaaugc	cuuucgcuuccuucgcuuuggu
7_40098	37.6	80	67	0	13	ucacgcuggcucccucucucugcag	ugugggagagggcugugcgggag
15_13560	37.5	72	68	1	3	cacccagaacugacagaccuagc	agguuuuucagcuucugguugc
5_33996	37.2	71	63	0	8	cagcacucagaggggcagaggga	uccugcccucccugggcug
10_465	36.3	76	61	0	15	gcucucgcucgcgggccgccucg	ggcgcggcuccgggucgucgagc
1_21565	35.4	69	62	0	7	ucuaagcagagguguuaguucc	aacugcaccucuacuuccaga
3_27907	34.9	59	45	0	14	ugucugcccccacccccacauc	cuggggguggggggggagacuac
11_4932	34.8	62	59	0	3	aacagaacugaaggacauacga	cucaugaccuucuggccuggguuu

15_13527	33.1	62	52	0	10	ucuggcacaggggugucuaggga	ucccacacacgccccuuugcccgc
11_4625	32.9	62	52	0	10	ugacugaaucuuguuaaagaau	ucuuuaacaagauucagucac
10_506	32.7	63	62	0	1	acuuaaccguuucuccggccu	ccuggaggagcgggagaga
2_26397	32.6	76	75	0	1	uuccaugauguagcuuguucugaug	auucuccagucaucuguggaau
1_23614	32.4	62	51	3	8	cugcauggcucugcauggcucu	uggcucugcauggcucugcaugg
1_23325	32.3	70	67	0	3	gcuaaacuuccuucugaugcuc	caacacaaggaaguuaggggcu
10_683	31.9	60	41	0	19	cagccggcagcuggccgucgcagc	gcuuuggccugcagccuggcugcc
6_37833	30.9	59	54	0	5	aagggaagcaagcucugcaugggu	uuugcagccugcaugcuacccucc
2_25671	30.8	58	33	0	25	acaggagcauagggcgcaggcagc	ugccuggcucccugcccucu
2_24390	30.7	59	43	0	16	gguuaagaucgccuuguuaggg	ucugacagugugauuuuaucug
2_24354	30.6	58	54	0	4	uuggaucugguguuguccacagacc	uguguggacacggcuccaug
10_11	30.2	64	52	0	12	cgcccgcuccggacgcccuuggc	cggggcuccggcgcgggaccgcu
11_5908	30.2	57	49	0	8	ugcucagaaacucuguccucccc	cgagacaguguagcugcggcag
14_10938	29.5	66	59	0	7	ucaacauggaagagcuggaugacu	cauugaacucuucugugugugc
5_34887	28.8	56	55	0	1	aguggaccggcagacaugucu	caucucuucggcccaaaucuaga
19_19955	28.8	55	54	0	1	aaguggacaggaauaaagagc	cuuugucucuguccucuc
15_12254	28.7	56	55	0	1	uccugagagagagaggaagu	cuucuucuauccucucaugcac
7_38567	28.6	61	54	0	7	uucccccgugcgggggggcggcggc	uccccgucccacucggcggg
6_37681	28.4	54	33	0	21	uuaggguuuucauugcugugaac	ucagagcaguagaaacucugacc
10_2431	28.3	63	59	1	3	uugggaguaugagaaacgaaauuc	auggguuucuccuguucccagg
2_24275	28	53	52	0	1	acgugaccucugucucccucagg	auggggaggcuggguguuauuugc
12_7126	28	54	50	0	4	uccucagcaugucuucagaacucug	aguucuaggaugcuuuggc
6_38061	27.4	53	22	2	29	gucacacucaguagagcagaga	uuugcuuacaugauugugcuca
18_18282	27.4	45	37	0	8	gaaagccagguaacaggugcugu	gguccugaaacuuggcuuugcc
11_4442	27.3	52	50	0	2	ucaagccagacccagugagcucu	ucucacugaggaugggcuuaagggc
14_11452	27.3	51	47	0	4	aucgccuccucagagaccuguu	aggucucugaggaggcgaucagu

6_37146	27.2	58	54	0	4	uucccccgugcgggggggcggcgg	cccgcccccacucggcggggc
7_39103	27	59	56	1	2	uucccuccgucuuccuuccagg	cgggaaggaaggaggcgggcuugacu
4_30900	26.8	51	43	0	8	cagagccuggucaaggucacuc	aguggucagggccaagccagaga
7_39418	26.6	50	39	0	11	uuaggcugaccaguacuggggcu	accaguacugggcuagccucuggccu
3_29829	26.5	62	56	0	6	ugguuuaucugaaauucugaacc	ucaaaauaucagaugcaccgcu
10_1016	26.3	58	55	0	3	ucuuggggaguugaugcucaga	uuggcuucuucuccgcaaacag
10_132	25.9	49	26	0	23	aaaaagaauuugcuccgaacu	uucggagcaaguuccuuuuca
11_3682	25.7	52	50	0	2	cagcccaucgacugcuguugcc	caacaucagucugauaagcuauc
8_42343	25.7	56	24	0	32	ccccgaagcuggucuaccucgagc	gucggguggcgcccgcguuggggu
14_10540	24.9	52	50	0	2	aaagaaaacucaggcugugga	acauccugagcguuccuccau
7_40082	24.7	41	39	0	2	ucugucccuucugccuuccagu	cugggaagggacuuaggacagac
11_4490	24.5	54	36	0	18	cgcuugaucuucucccucugcagg	gugcggaaggacagauccugug
15_12311	23.9	44	40	2	2	ugcgcgaguugaggcgggucuau	cuccccgcugggccucgc
2_23981	23.6	54	52	1	1	uauucugacucuaccuaacugga	ucugcuagguagugguggaacu
18_18957	23.6	46	43	0	3	uuuugagacaggcugaccugggau	ccaggcugguuggucuucaacu
8_42685	22.9	43	25	0	18	ccaacucugcauucuagcugcc	caggggguuugcaggauuggaga
4_30826	22.8	43	41	1	1	ucuuuuagaaggcguccaguaga	ucuggaggccuucuuuuugg
15_13669	22.8	43	40	0	3	uagacacuacuggccuucccagga	cugauuggcuguggguccau
11_5759	22.2	50	47	0	3	uguggaggaggagccugaggcu	ucucacacacuuccuccagg
15_13448	21.9	41	40	0	1	augcuuccaccucaugggcgggc	cagcccaugggguggagcugcc
8_42518	21.6	49	32	0	17	ugacugguguuccuuccgcaga	ugcuggacaggugccauggacccu
9_45310	21.5	40	29	0	11	caaggauguucgcgcggcuucu	ccggccgccggaucccccuuguc
9_45230	21.2	40	39	0	1	cuccagcauuguuccucugcaaa	ccagaggaaaaguugcaggccuug
7_40712	21.2	48	24	0	24	auaccucacuucuacucuggcag	uccaguagagaaguggcauga
11_5310	21	49	47	0	2	aaagaggaggggguggac	uaaccucuccuguagccc
8_42029	20.8	39	37	0	2	ucccagacuuccucaggcuuuc	agcucuguggaaguuggggugugu

5_34069	20.8	46	37	0	9	aucccccccgcaccccaacagg	ucgggguggggguggggucugccc
11_3379	19.9	37	34	0	3	ugcguggcacuguggcuggggggaa	cuuccagaguggacaggc
11_5814	19.1	45	43	0	2	ucccaucccgcucuuggcucagg	aggccuauggcuggguggcugggcuu
3_28429	19.1	37	35	0	2	acagcuccucucucucugaag	ugcagagagaugcggggaggguuaa
19_19217	19	38	37	0	1	uaaggaggaguaaugaguucuc	aaacucauuaaccccccugaga
8_42228	18.7	33	32	0	1	uucccaagucggaugcugcgcc	gcagcauccgacuugggaac
X_46390	18.6	47	41	0	6	aucucggccuuccagucucugc	ugggcuuggacucggccggcaaga
15_13683	18.4	34	18	0	16	agaaggaacaguggcgccaacag	uguggggccagggugacuucagc
19_20436	18.2	37	24	0	13	uugagacugaaguucacuucu	uagugaaccaugaacucagca
18_18115	18.1	34	29	0	5	aggagggccuagaauucauagag	uguguauuccugguuguucug
11_5986	17.8	33	18	0	15	caaggcucugggcaugggggacu	ucuccaaagccaagggcacuguug
13_9041	17.7	34	31	0	3	acgggauuuaaagauaacuca	gacuuaucuuaaagccuguca
8_41566	17.6	32	24	0	8	acugagcaccggccgcacgcgc	ccgcgcgaccgggacucaga
3_29609	17.5	42	37	0	5	aaggcuacuaagcaaugguug	aaccauggcacugguguagccca
1_23413	17.2	34	33	0	1	caggaaccguuuucucuaugacu	ugaugggguaaagugguucuu
3_29689	16.9	39	36	0	3	aggggugguggcaucugccugagga	ucgcagcauccacacccucac
17_16966	16.8	38	36	0	2	uccccgcgagcgccgagcccu	ggcccggaccuggcuggggggccg
12_6778	16.8	31	27	0	4	augcaugcgcacaguaggacc	accacuguggugucuguaugua
15_12217	16.8	33	22	0	11	ugggacagaauccaacuaugaa	acaguuagacucugucccauggag
7_39578	16.5	31	24	0	7	agagggccuccacuuugauggcc	cacacaaaguggaagcacuuucu
14_11007	16.4	30	29	0	1	agagucucuggagagagu	ucucucugugucucugug
16_14520	16.2	30	13	0	17	aagccaggcccagugucacuc	uaugucacugggcauuggcuuuga
15_13871	15.9	30	20	0	10	ugggagauacugcuaagaugc	cucuguaccagcaucucccu
17_16684	15.9	30	29	0	1	ggaaggugggugcuaagggcuga	aguucauagcaaccccccuuaagccug
12_8209	15.7	37	30	0	7	ugcagcggcucaccggccugc	cagugccugcugagagcugccuc
7_40811	15.4	37	31	0	6	uggggcuaaaggucugugccacu	uggccaggauuuuaguccugau

14_10921	15.2	27	24	0	3	cucgcugcacaucugccacugcu	cggcggcgacggcagcgagcc
4_30830	15.1	28	27	0	1	uuugcucuggcuugacuggcc	ccagaggacccagagccugcu
4_31354	15	28	25	0	3	acuggcggcgguugcucucugc	cagagacuaaucugucgccacc
4_30911	15	28	15	0	13	uguaguuucaguuuuccuaacu	aguuaggaaaacugaaacu
2_25240	14.9	37	27	0	10	agggcuggucaacaaguaggaagg	uuccuccuugguccugccccaga
15_12848	14.9	27	16	0	11	aggagaaggguaggcugcuauu	uaguggccaccuugaucugau
7_39887	14.8	20	18	0	2	gagaacuucuaguguccccugcu	agagaccuggcugucucug
16_14250	14.7	28	25	0	3	ccucucuccaguauagcuuuc	aagcuacaacugaaggguagagcc
19_20494	14.7	27	14	0	13	cugaggccuguggcaccacau	gaggugcccuuggacucauga
11_5771	14.2	27	25	0	2	aucucaccuagacucuucugcu	caggugaaggucgagaugagacau
7_38830	14	32	16	0	16	gcgggggugcgugccgcgagg	cccgcggcggccgcccccguccu
6_37184	13.6	24	17	2	5	aaggcaaaucaucucucuggcu	cagagagaugacuugccuugua
X_47332	13.5	25	18	0	7	acucgaaaaguuuuggauuuugg	aggucuaaaacauuuugagucu
10_1872	13.5	32	31	0	1	ugacccggccucucccccccagg	cugggccuuggucccuggugagug
11_2537	13.4	24	13	0	11	ccacucugcugggcagcccag	uggcugccagcagagccugga
3_28805	13.4	27	20	2	5	uaguaugauguccaauuuucugagc	cagaagauccuggacaga
8_43188	13.2	25	19	0	6	agacaccgugacuaagacaacc	cugucuuaguuaggguuuuacugcu
6_37711	13.1	32	30	0	2	ucuccugagugcugcagcuggca	ccaguucugcacagcccu
5_35460	13	24	19	0	5	cggguuacauagcaggcugacu	ugugagucuguuaugugcccagg
10_1837	12.8	37	27	0	10	ugagaggcacucugguuuuguga	cuagccaggcugaugucucucuaacc
12_6773	12.8	25	24	0	1	ugcugaauccagagguuacaucu	ggugaucucggagauucaggugu
17_15765	12.8	23	19	3	1	aguccugguguuggcugggagc	acuuggcuagcaccagggacacgccu
13_9210	12.7	24	20	0	4	uccggggauguuuuuguugcuc	agccgcuaagaacuucccccaagc
15_13271	12.6	23	20	0	3	uggccugagagguugugaugccu	ccacacggccagucaggcag
19_20193	12.6	17	15	0	2	aaggcucugugagauuugcaaug	cugcaaaguagcagagcuguca
19_19782	12.1	27	14	0	13	ucucuucaacuguucugguagg	aaccaaacagcugugaaaagga

9_44761	12	14	13	0	1	aaugugacucagcuaccugaac	ucagacugcugagucacauugca
7_39386	11.7	23	22	0	1	ugugagacuuuguagcagcgu	gaaguuacucugucucacuugu
17_16932	11.7	29	24	0	5	acagcugcagacaguggccccgu	uggccucucccugaagcucug
11_3843	11.6	22	20	0	2	uacuggccacugcucaggggaca	gccccgggcucugguugcaguucc
1_23682	11.5	21	18	0	3	cacauggaguugcuguuacacc	gaaguaacagcagcuccacugg
8_41495	11.2	20	10	0	10	aaggcagugccuccagcgggc	uggcuguaggcagugccu
7_40714	10.8	31	29	0	2	ucuguugcucccccucugccaaca	augggguggggaguaaggaugagaag
3_27935	10.5	25	22	0	3	ugauugcuguauggcuuaauuc	auaagccaccauagcaagcacug
8_40985	10.4	21	15	0	6	ucacauccuugggauucugccu	gaaggaucccaaaggcuugauug
3_28881	10.3	20	18	0	2	agcaggcaggguaacuuuagagc	ucuauuguugccugcuuggccaaa
11_3101	10.3	19	10	0	9	cucccuggucuuaccauuacc	uagugguacuagcucugggugcc